UNIVERSITY OF LEICESTER

Evolution of Plant Male Germline-Specific Transcription Factor DUO POLLEN 1

Thesis submitted for the degree of

Doctor of Philosophy

By Mingmin Zhao BSc (China)

Department of Genetics

University of Leicester

September 2017

Evolution of Plant Male Germline-Specific Transcription Factor DUO POLLEN 1

Mingmin Zhao

Abstract

Flowering plants account for the 30 crops that provide 95 % of the food for humans. The reproduction of this group depends on the production of two twin sperms. The establishment of the male germline lineage requires the transcription factor DUO POLLEN 1 (DUO1). DUO1 is required for both the cell cycle progression and sperm cell differentiation. This thesis focused on the origin of DUO1 and its target regulation.

Much work was dedicated in searching the evolutionary origin of DUO1 in the R2R3 MYB clade. Based on the analysis of sequences homologous to DUO1 and its sister clade GAMYB, the earliest DUO1 homolog was identified in the green algae. The DUO1 clade did not proliferate after multiple polyploidy events, possibly restricted by its male germline-specific role supported by transcriptome data. The ancestral DUO1 experienced a major MYB domain sequence change in the bryophytes and a second change in the C-terminus in the angiosperms. The MYB domain changes caused a change in the target DNA sequence, which has then been conserved among Embryophyta DUO1 homologs. Another change also happened in the region where a miR159 binding site is present in most angiosperm *DUO1* homologs. Sequence and functional analysis showed that this change evolved long before the emergence of miR159. The changes in the C-terminus of DUO1 led to a higher target promoter activation capability in the angiosperm homologs, which was confirmed by functional tests of the angiosperm and bryophyte DUO1. This C-terminal region contains the transactivation domain (TAD) of DUO1 and certain functionally important motifs were highlighted in the study. While these motifs indicated that DUO1 was a member of a TAD family, it was also demonstrated that unknown sequences carry critical features for activation. Together these results mapped the evolution history of DUO1 in the Streptophyta lineage.

Acknowledgement

It is a pleasure to thank the many people who have made this thesis possible.

First and foremost I would like to thank my supervisor Professor David Twell. His guidance, support, and encouragement have helped through my PhD and will continue to influence my future career. His mentoring pushed me to overcome the difficulties in my early years of research, and the long discussions with him have enlighten me to think critically and thoroughly. I am proud to have worked with him and to be involved in some great science.

I would also like to thank Dr Nicholas Rutley for his enthusiasm, advice, and friendship in and outside the lab.

Additionally I would like to thank past and present members of the Twell lab group for their help and support: Dr Michael Borg, Dr Dieter Hackenberg, Dr Lewis Collins, Dr Ugur Sari, Dr Nadia Taimur, Dr Abdur Rauf, Dr Huijuan Yang, Dr Yanfeng Kuang, Jaypal Darbar, Yosra Al-Hakeem, Ghazwan Hasan, Ramesh Patel, and Sanjeev Kumar. For all the undergraduate students who have worked their projects with me, it has been a pleasure to be involved. I would also like to thank other people from the Department of Genetics who has offered me help. Furthermore, I would like to thank the people from the Department of Molecular and Cell Biology for their assistance, including but not limited to: Dr Ralf Schimid for his guidance and advice on protein structures; Prof John Schwabe and Dr Peter Watson for the collaboration in the mammalian two-hybrid assays; Dr Xiaowen Yang, who is sadly no longer with us, for his technical help and example as a great yet humble scientist.

Special thanks to all my collaborators from all over the world for the great work we have done together.

Friendships have been instrumental for helping me through the long days of the PhD life. My housemates and friends from the Taekwondo club, Erasmus program, first aid society, and elsewhere have made life in a foreign country a lot easier.

I would like to reserve special thanks to my love Amy Zsinkó, who has made me feel home wherever we are together.

I would not be where I am today without the support of my family. I dedicate this thesis to my parents along with my endless gratitude and love. My father gave constant guidance to me in science and life in my young age. My mother has always been encouraging me to aim and achieve higher. My grandparents took care of me since I was born and saw what I could achieve before I could even believe it. My cousins who were the perfect siblings I never had, gave me wonderful memories that helped me through dark times.

“There is a grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.” *On the Origin of Species*

- *Charles Darwin*

Table of Contents

List of Figures

List of Tables

Abbreviations

aa	amino acid
ANOVA	analysis of variance
ATP	adenosine triphosphate
BAP	6-benzylaminopurine
BASTA	glufosinate ammonium
BCP	bicellular pollen
bp	base pair
BSA	bovine serum albumin
β-gal	β-galactosidase
°C	degrees centigrade
CDKA;1	Cyclin-dependent kinase A;1
cDNA	complementary DNA
CDS	coding sequence
χ^2	Chi-square statistic
cm	centimetre
CoA	coenzyme A
Col-0	Columbia-0
CYCB1;1	CyclinB 1;1
DAPI	4’,6-diamidino-2 phenylindole dihydrochloride
DAA1	DUO1-activated ATPase1
DAZ1	DUO1-activated zinc finger 1

DAZ2	DUO1-activated zinc finger 2
DAZ3	DUO1-activated zinc finger 3
DMEM	Dulbecco's modified Eagle's medium
dH_2O	distilled water
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
DTT	dithiothreitol
DUO1	Duo Pollen 1
DUO3	Duo Pollen 3
EDTA	ethylenediaminetetraacetic acid
F1	first filial generation
FACS	Fluorescence-activated cell sorting
FBL17	F-box-Like 17
FLuc	firefly luciferase acitivty
FLuc/RLuc	normalised dual luciferase activity
fmol	fentamole
g	gram
g	gravity
GAMYB	gibberellin-and abscisic acid-regulated MYB / GAMYB-like family
GCS1 / HAP2	Generative cell-specific-1 / hapless-2
gDNA	genomic DNA
GEX2	Gamete-expressed protein 2
GFP	green fluorescent protein
GUS	β-glucuronidase

H2B	histone H2B
HEK 293	human embryonic kidney cells 293
HTR10	Histone3-related 10 variant
kb	kilobase pair
KRP	Kip-related protein
l	litre
LB media	Luria Bertani media
LUC	firefly luciferase protein
M	molar
mCherry	monomeric cherry
MES	2-(N-Morpholino)ethanesulfonic acid
mg	milligram
MGH3	male gamete-specific histone-3 (HTR10)
miR159	microRNA159
ml	millilitre
mM	millimolar
MP / MPG	mature pollen / mature pollen grain
mRNA	messenger RNA
MS	Murashige and Skoog
MS1	Male Sterility 1
MYB	myeloblastosis family transcription factor
No-0	Nossen-0
ng	nanogram
OD	optical density

p	plasmid
PBS	phosphate buffered saline
PEI	polyethylenimine
PCR	polymerase chain reaction
PHD	plant homeodomain transcription factor
PMI	pollen mitosis I
PMII	pollen mitosis II
PPT	phosphinothricin (glufosinate)
prom	promoter
qRT-PCR	quantitative RT-PCR
RenLUC	Renilla luciferase protein
Rluc	Renilla luciferase activity
RFP	red fluorescent protein
RMS	root-mean-square
RNA	ribonucleic acid
ROI	region of interest
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcriptase-PCR
SE	standard error
SNP	single-nucleotide polymorphism
TAD	transactivation domain
TAE	tris-acetate EDTA
TCP	tricellular pollen

TCP	TCP protein domain family transcription factor
TF	transcription factor
Tm	melting temperature
Tris	tris(hydroxymethyl)aminomethane
Triton X-100	t-Octylphenoxypolyethoxyethanol
U	enzyme units
µg	microgram
µl	microlitre
µM	micromolar
UNM	uninucleate microspore
UTR	untranslated region
V	volts
v/v	volume per volume
WGD	whole genome duplication
w/v	weight per volume
WT	wild type

Nucleobases:

A	adenine
C	cytosine
G	guanine
T	thymine
U	uracil

Amino acids:

A/Ala	alanine

R/Arg	arginine
N / Asn	asparagine
D / Asp	aspartic acid
C / Cys	cysteine
Q / Gln	glutamine
E / Glu	glutamic acid
G / Gly	glycine
H / His	histidine
I / Ile	isoleucine
L / Leu	leucine
K / Lys	lysine
M / Met	methionine
F / Phe	phenylalanine
P / Pro	proline
S / Ser	serine
T / Thr	threonine
W / Trp	tryptophan
Y / Tyr	tyrosine
V / Val	valine

Species abbreviations: binomial nomenclatures were abbreviated to either the 2-letter initials or 3-letter including the second letter of the species name (mainly for phylogeny), e.g.

At / Ath *Arabidopsis thaliana*

Chapter 1 Literature Review

1.1 Plant male germline and double fertilisation

The flowering plants, or angiosperms, make up around 90 % of all plants species on earth today. These plants also account for about 95 % of the food source for humans. Angiosperms, like all other land plants, experience the alternation of generations in their life cycles, in which a sexual haploid gametophyte (n) alternates with an asexual diploid sporophyte (2n). The sporophyte produces spores (n) through meiosis, which then grow into a haploid gametophyte (n).

In animals, the separation of the reproductive cells and somatic cells happens in the early embryo stage (Strome and Lehmann, 2007). On the contrary, the stem cells of the flowering plants remain undifferentiated in the meristems and germline cell fate is determined in the stamen and ovary. Meiosis marks the transition from the sporophytic generation to the gametophytic generation (Bhatt et al., 2001, Wilson and Yang, 2004).

Both male and female gametophytes (MG & FG) will continue the differentiation process by going through more cell divisions. The FG usually develop in to multicellular structures with two cells, the egg cell and the central cell each of which fuse with a single sperm cell upon fertilisation. In the process of FG development, three out of the four initially produced megaspores undergo programmed cell death (Yadegari and Drews, 2004, Yang et al., 2010).

For the male lineages, each microspore will go through an asymmetric division, which is the establishment of the germline (Twell et al., 1998). Then the reproductive cell has to finish a round of mitosis before the pollen grain is finally mature. In contrary to the female megaspores, no microspore is lost during the development (Twell, 2011). The two sperm cells will fertilize the egg cell and the central cell, producing the embryo and the endosperm, respectively. This process of double fertilization greatly benefits the seed production (Walbot and Evans, 2003, Berger and Twell, 2011). It is believed that the nutritious pollen grains were also responsible for the angiosperm-pollinator coevolution, which contributed to the diversity of this clade (Lunau, 2004, Hu et al., 2008).

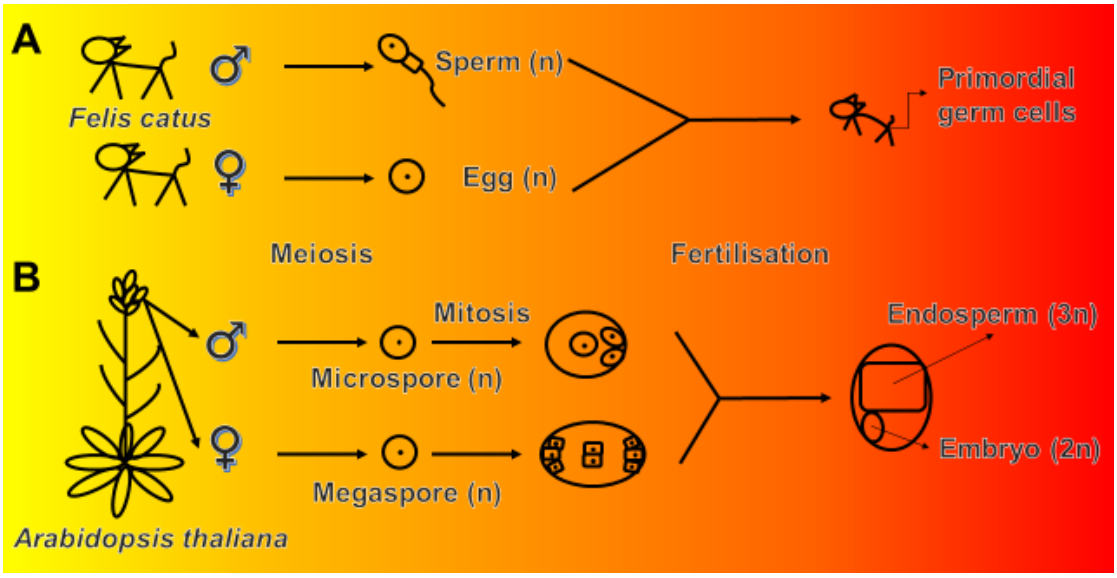


Figure 1. 1 Differences in animal and plant reproduction.
The animal germlines have their cell fate decided early in embryogenesis, and undergo meiosis later in life. Flowering plants establish a true germline only in a reproductive phase, and not until the two mitosis after the meiosis is complete. The two differentiated sperm cells in a pollen grain would eventually fuse with the female egg cell and central cell to produce an embryo and an endosperm. Diagram based on Walbot and Evans, 2003.

Despite having a highly reduced haploid gametophyte generation, there are numerous genes and networks involved in the flowering plants male germline development (Borg et al., 2009). A pollen transcriptome of *Arabidopsis thaliana* showed that 40 % of the mRNAs detected were pollen-specific (Honys and Twell, 2003). The development of the MG is accompanied by the magnitude increase of the MG-specific transcript ratio (Honys and Twell, 2004).

Recent advances in the field have made a number of tools becoming available, such as Plant Male Reproduction Database PMRD (Cui et al., 2012), FlowerNet (Pearce et al., 2015), or more comprehensive databases that includes all types of eukaryotes, like MeioBase (Li et al., 2014a), Tree of Sex (Ashman et al., 2014). These tools have become very helpful in understanding the relationships between reproductive genes and developmental processes in flowering plants. For example, microarray data were used in the recent discovery that the *ABORTED MICROSPORES* (*AMS*) is required for early and late pollen formation (Ferguson et al., 2017).

1.2 DUO1 and male germline development

DUO POLLEN 1 (DUO1) is a key regulator of male germline development in plants. It is required for both sperm cell division and differentiation (Durbarry et al., 2005, Rotman et al., 2005, Borg et al., 2011). Studies of some direct DUO1 target genes, including the DUO1 activated zinc finger (DAZ) proteins *DAZ1*, *DAZ2*, *DAZ3* and *DAZ3-like*, and the male germline-specific histone H3 variant (H3.10) termed *HTR10*, suggest they form a network controlling the male germline cells development (Brownfield et al., 2009a, Borg et al., 2011, Borg et al., 2014).

Little was known about *DUO1* when it was first sequenced in the year 2000 with the whole *Arabidopsis thaliana* chromosome 3 (Kaul et al., 2000). It was labelled as *MYB125* due to its R2R3 MYB domain. The *DUO1* gene accession number is GI: 18411615 and the TAIR is At3g60460. In the first phylogenetic study of the MYB family proteins that included the *DUO1* locus (identified as At3g60460), it was grouped with a clade of other R2R3 MYB proteins known as the GAMYB family (Jiang et al., 2004). It was pointed out that *DUO1* and all of the *GAMYB* family genes possess a potential miR159 binding site.

In 2005, two classes of *A. thaliana* mutants affecting pollen development were described that contain only one reproductive cell, instead of two (Durbarry et al., 2005). They were named *duo pollen 1* (*duo1*) and *duo pollen 2* (*duo2*) as germ cell division was blocked, leaving the mature pollen grains with only a vegetative cell and a single undivided generative cell rather than two sperm cells. The study also showed that the *duo1-1* mutant pollen failed to go through the G2-M transition (Durbarry et al., 2005). In the same year, an additional allele *duo1-2* was identified and the mutations responsible for these loss of function alleles were described (Rotman et al., 2005). It identified DUO1 as a R2R3 MYB transcription factor (TF) containing a supernumerary lysine residue compared to other plant MYB sequences at position 58, which is labelled as the Lysine 66 due to its position in the *A. thaliana* DUO1 sequence (Rotman et al., 2005). In this study a previously identified tobacco protein known as B25 (Kyo et al., 2003), a rice putative protein, and three maize predicted proteins were also considered DUO1 homologs, identifying apparent DUO1 homologs in dicots and monocots. More DUO1 homologs were identified later in *Physcomitrella patens* and *Selaginella moellendorffii* (Brownfield et al., 2009a). The pre-angiosperm sequences however were incomplete, poorly

annotated at the time and not consistent with the latest prediction of the sequences, as the C-terminal regions were either missing or incorrect.

In the following studies, several important genes were found to be regulated by DUO1 such as *GCS1*, *GEX*, and *CYCB1;1*, which were later regarded as direct targets based on the DUO1 binding site (Borg et al., 2011). With more genes identified as being controlled by DUO1, a male germline regulatory network model was established, with DUO1 being the master regulator to control other key germline specific transcription factors such as DAZ1 and DAZ2. (Borg et al., 2011, Borg et al., 2014).

Brownfield et al., 2009 described the roles of DUO1 in promoting cell differentiation as well as in cell division, which is very unusual for a single regulator. There were previously known examples of one gene that regulates both processes, like FAMA in *A. thaliana* stomatal development that has been shown to coordinate both processes by promoting differentiation and halting cell division (Ohashi-Ito and Bergmann, 2006), and Pax-3 in vertebrates that promotes the cell cycle through locking cells in an undifferentiated state (Lang et al., 2005, Doddrell et al., 2012). A rare example of Cyclin D1-3 in human performs similar positive regulations on both processes like DUO1 (Pauklin et al., 2016). It is required for the G1-S progression in the cell cycle, and controls cell fate decisions in human pluripotent stem cells. The plant male reproduction related PTC1, a PHD transcription factor involved in the GAMYB pathway (Aya et al., 2009), controls the tapetal cell apoptosis (a process closely linked to cell cycle (Pucci et al., 2000)) and pollen formation in *Oryza sativa* (Li et al., 2011a). These processes also require the involvement of MYB80/103 (Phan et al., 2011, Phan et al., 2012, Xu et al., 2014b), a downstream regulator of AMS (Lou et al., 2014, Ferguson et al., 2017).

Understanding of the cell cycle in the plant male germline has been established over the years (Figure 1. 2 A). Fundamentally speaking, just like other eukaryotic organisms, plant cell cycle is controlled by the cyclin-dependant kinases (CDKs). Specifically, CDKA is vital for both G1-S and G2-M transitions (Inze and De Veylder, 2006). In *A. thaliana*, only one homolog CDKA;1 exists, and is essential for the pollen mitosis II (PM II) that produces two sperms cells. There is only one sperm-like cell, although fertile *per se*, in the *cdka;1* mutant and it preferentially fuses with the egg cell (Iwakawa et al., 2006). There are two known Kip-related proteins (KRPs) that inhibit CDKA;1 in the male germline, KRP6, and KRP7, and they are targeted by an F-box protein FBL17 that forms

an SKP1-Cullin1-F-box protein (SCF) E3 ubiquitin ligase complex (SCF^{FBL17}) (Verkest et al., 2005, Kim et al., 2008a). Similarly, DUO POLLEN 1 (DUO1) and DUO POLLEN 3 (DUO3) are also essential for the G2-M transition during pollen mitosis II in *A. thaliana* (Brownfield et al., 2009a, Brownfield et al., 2009b). For DUO1, the cell cycle progression function is DAZ1/2-dependant (Borg et al., 2014).

In contrast, the cell differentiation processes in male germline are less well studied. Currently many processes and genes have been identified as being involved in pollen development. For example, the pollen-specific *LAT52* is required for pollen hydration and pollen tube growth (Twell et al., 1990, Twell et al., 1991, Tang et al., 2002). The tapetum-specific *AMS* is associated with tapetal function and pollen wall formation (Xu et al., 2010, Xu et al., 2014a, Ferguson et al., 2017). The more intriguing genes in our understanding of the germline cell fate decisions are the male germline-specific ones, like the generative cell-specific protein gene *GCS1* (*HAP2*) which is required for pollen tube guidance and gamete fusion (von Besser et al., 2006, Wong and Johnson, 2010). Another such case is the *GAMETE EXPRESS 2* (*GEX2*), which is essential for the gamete attachment in *A. thaliana* (Mori et al., 2014). For the histone H3 variant H3.10 (or *HTR10* for the gene), its specific function is yet to be discovered (Okada et al., 2005, Borg and Berger, 2015). Interestingly, although DAZ1/2 alone cannot properly facilitate the cell differentiation, the presence of at least one is still required (Borg et al., 2014). Unlike the aforementioned *cdka;1* mutant, the *daz* double mutant sperm-like cell is not fertile. This makes DAZ1/2 transcription factors required for both cell division and differentiation, just like DUO1.

In the absence of DUO1 (Figure 1. 2 C), although the microspore still goes through the asymmetric division during the pollen mitosis I, the bicellular pollen cannot mature into tricellular pollen. Unlike the sperm-like cell in *cdka;1*, the germ cell in *duo1* is not fertile, nor does it express the male germline-specific histone H3 variant *HTR10* marker.

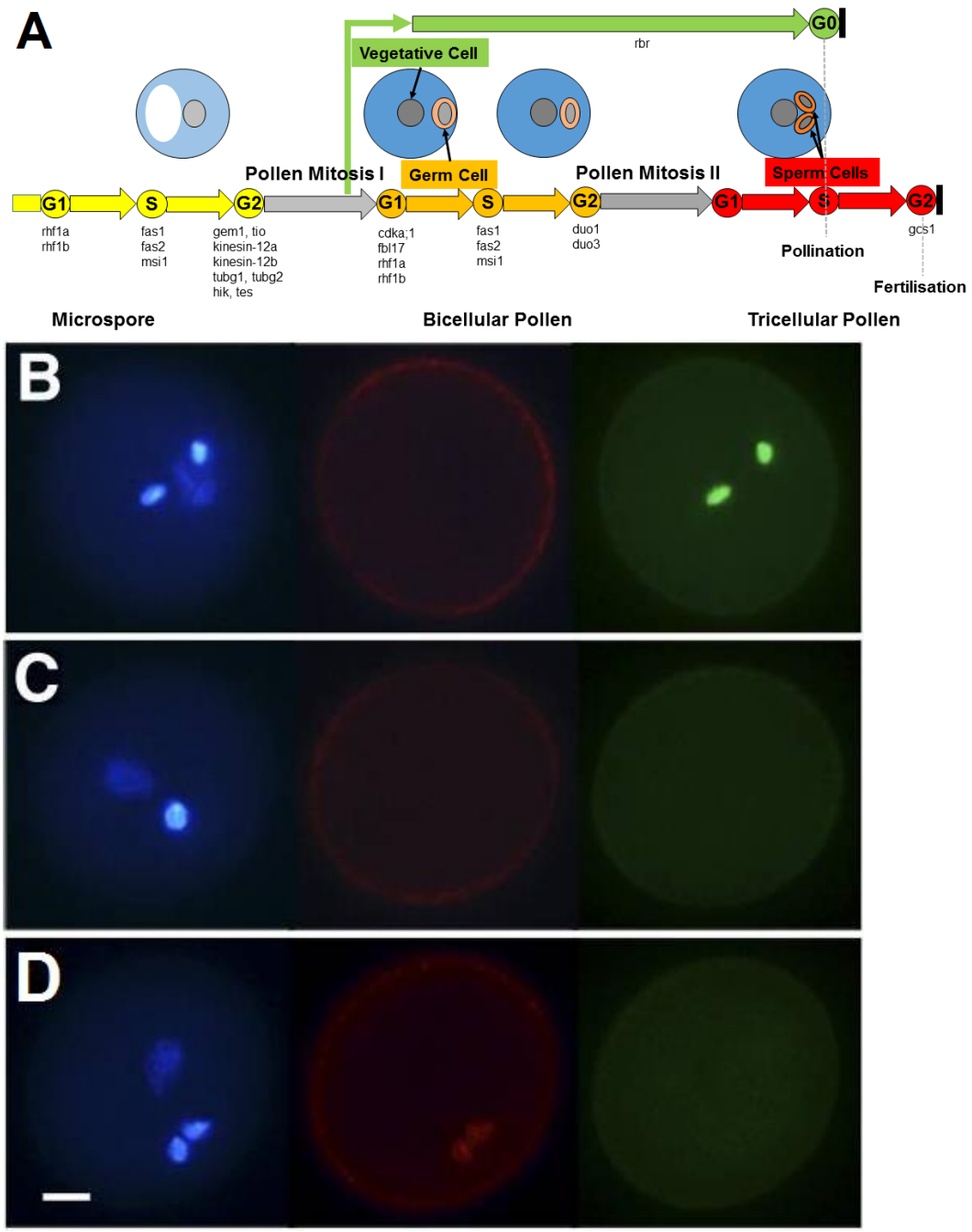


Figure 1. 2 DUO1 is critical for G2-M cell cycle progression and sperm cell maturation. A. Male germline development and cell cycle schematics adapted from Twell, 2011. Cell cycle progression regulators are mapped in the context of gametophyte mutations (Berger and Twell, 2011, Twell, 2011). In *duo1* mutants, the germ cell fail to divide and develop into two sperm cells. B - D. The functions of DUO1 are typically measured by the ability to promote cell division and differentiation. The pollen grains were captured under the florescent microscopy by (Borg et al., 2014). The GFP was attached to the nucleus locating histone H2B, and driven by the DUO1 direct target and germline-specific *HTR10* promoter to mark cell differentiation. Transgenes used for complementation studies were labelled with certain form of red florescence protein (RFP), such as mCherry. (B) WT tricellular pollen (TCP) with GFP signal. (C) duo1-1 bicelluar pollen (BCP, cell cycle defect) with no GFP signal (cell development defect). (D) Expression of DAZ1 in duo1-1 mutant rescues cell division (TCP), but not cell differentiation indicated by the lack of GFP expression.

After the discovery of DUO1 functioning as a regulator for cell division and differentiation in *A. thaliana*, it has been suggested that these two processes are regulated by a conserved DUO1 network. This conservation was later extended to a range of anther and pollen developmental pathways (Gomez et al., 2015). The expression patterns for DUO1 homologs in other species also suggest that they all have conserved functions in these two aspects, at least within angiosperms where cell cycle progression is required for the male germline development process (see Chapter 3).

This idea can be confirmed by complementing the *duo1* phenotypes in *A. thaliana* using the aforementioned DUO1 homologs. Much of these functional complementation analyses were performed by Dr Ugur Sari (Sari, 2015). The details of the rice and tomato DUO1 homologs complementation experiments are described here as an example of how functional tests are typically performed. The rice *OsDUO1* cDNA was amplified from anther RNA from *Oryza sativa japonica* 9522 as described in Li *et al*., (2010). Similarly, the tomato (*Solanum lycopersicum*) *SlDUO1A* and *SlDUO1B* cDNA was prepared as described by Sari, 2015. The OsDUO1 complementation was done using the *duo1-1* GFP marker line detailed in Chapter 2, while SlDUO1A/B was tested on the heterozygous *duo1-4* lines thus the result does not include transactivation ability.

Sari, 2015 reported that both the rice and one of the tomato (A) *DUO1* homologues were able to rescue the failure of germ cell division and differentiation in *duo1-1* pollen in *A. thaliana*. For promDUO1:AtDUO1-mCherry and promDUO1:OsDUO1-mCherry (a form of RFP), germ cell division rescue was determined by scoring the increase in the percentage of tricellular pollen, which can vary from 50 % (no rescue) to 75 % (full rescue). Gamete differentiation was evaluated by calculating the proportion of cells which express the germ cell-specific marker, promHTR10:H2B-GFP, and by determining the transmission efficiency of the *duo1-1* allele based upon antibiotic resistance of seedlings resulting from crosses to homozygous male sterile *ms1* pistils. For transactivation, heterozygous *duo1-1* plants showed no significant deviation from the predicted values of 50 % TCP and 50 % GFP positive pollen.

In Sari, 2015 five independent single insertion T1 lines of promDUO1:AtDUO1-mCherry showed no significant deviation from the theoretical value of 75 %. Two independent single insertion T1 lines (~50% RFP+) of promDUO1:OsDUO1-mCherry were screened and scored in detail. Interestingly, they also showed no significant

deviation from the theoretical value of 75 %, suggesting that OsDUO1 was as efficient as AtDUO1. SlDUO1A also demonstrated a full cell cycle defect rescue with the TCP:BCP=732:280 (2.6:1, not significant from 3:1, significant from 1:1) in eight single insertion T1 lines. However, due to the lack of expression for SlDUO1B, we do not have any evidence for its ability *in planta*. This phenomenon is not unique to just SlDUO1B though. Both DUO1 homologs from *Physcomitrella patens* (PpDUO1A and PpDUO1B) and *Amborella trichopoda* failed to express in the pollen. This is possibly due to protein instability in the artificial environments. The ability to rescue *duo1-1* male transmission was evaluated using two T1 lines for promDUO1:AtDUO1-mCherry and promDUO1:OsDUO1-mCherry, and three lines for promDUO1:SlDUO1A-mCherry. The percentages of PPT resistant seedling observed did not deviate significantly from the theoretical maximum value of 67 % (Figure 1. 3).

The dual-luciferase transient assays provided a quantified level of target promoter activation and had been used in many cases to determine the functional abilities for the angiosperm DUO1 homologs (Figure 1. 4).

Note that the tools used in the research described as "promoters" are the regions of around 1000 bp upstream of the start codon ATG. These regions should be considered to include all types of cis-regulatory elements like enhancers or insulators. Unless specified, all promoters or "prom" in the names of constructs refer to these regions.

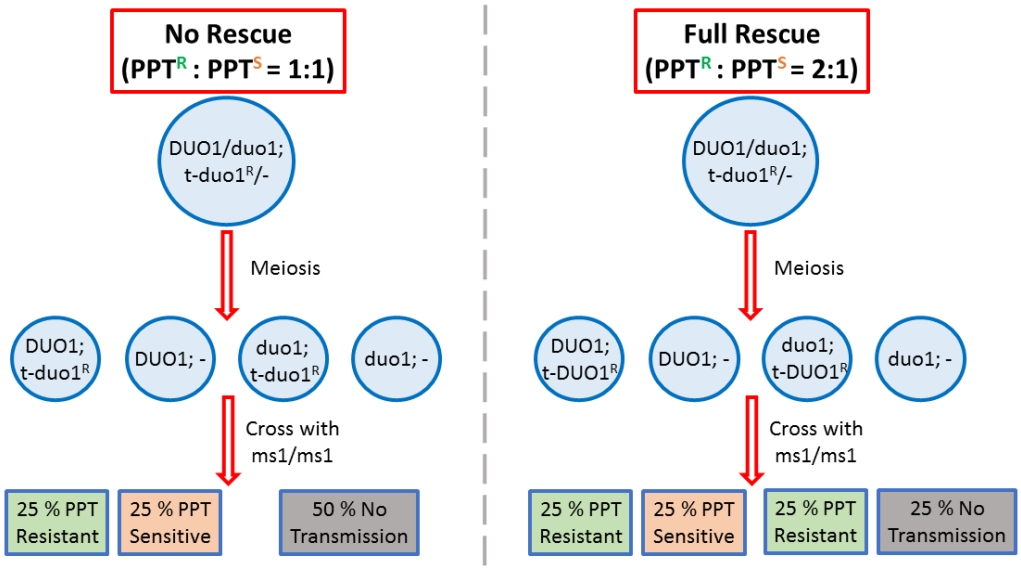


Figure 1. 3 Diagram of transmission measurement using PPT resistance.
The gametophytic transmission nature of DUO1 ensures that the fertility of 25 % of the gametes are dependent on the transgene, causing a range of antibiotic resistant seedling ratios.

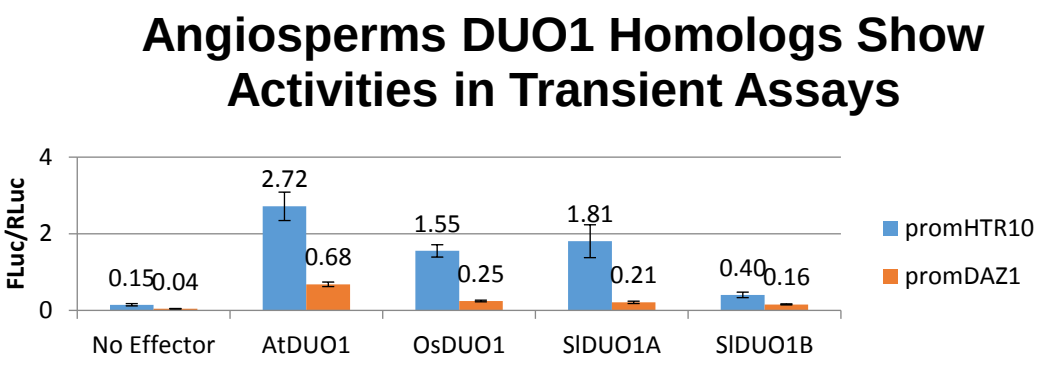


Figure 1. 4 Examples of DUO1 homologs *in vivo* test using known *Arabidopsis thaliana* DUO1 direct target promoters in tobacco leaves.
Rice and tomato DUO1 homologs have been tested on both the *HTR10* and *DAZ1* promoters. The promHTR10 is more sensitive and is ideal for transactivation measurement, although promDAZ1 shows very similar trend at a lower level. Raw reinterpreted from Sari, 2015.

1.3 The MYB protein family

DUO1 belongs to one of the most abundant family in plants, the MYB family (Rotman et al., 2005). In 1982, the first MYB gene *v-myb*, an avian myeloblastosis virus oncogene, was sequenced (Klempnauer et al., 1982). Other MYB family members like A-myb, B-myb and C-myb were also found in vertebrates and proven to regulate proliferation, differentiation, and apoptosis (Weston, 1998, Oh and Reddy, 1999, Beall et al., 2002). The first MYB gene identified in plants was the c1 locus of *Zea mays*, also known as *ZmMYBC1*, which was thought to act as a transcription factor (Paz-Ares et al., 1987, Cone et al., 1993). Despite the highly diversified sequences of the MYB family genes (Ito, 2005), they all share a conserved DNA-binding domain (Peters et al., 1987), known as the MYB domain. The MYB domain consists of up to three imperfectly conserved repeats (R1, R2 and R3), which usually have 51 to 53 amino acids. The regularly spaced tryptophan residues within each repeat, normally three separated by 18 to 19 amino acids, is the most distinctive feature of all the MYB proteins (Saikumar et al., 1990). These tryptophan residues, flanked by basic amino acids, are essential for maintaining the helix-turn-helix structure of the DNA binding domain. They form a hydrophobic core and arrange the adjacent amino acids in the appropriate place to interact with the sequence-specific target DNA (Saikumar et al., 1990, Heim et al., 2003).

There is a huge difference in the MYB protein numbers between plants and animals. So far only a few MYB proteins were identified in animals. It seems there is only one MYB transcription factor in most invertebrates and three in vertebrates (Lipsick, 1996,

Andrejka et al., 2011). On the contrary, the numbers of MYB-related proteins in plants are highly abundant due to gene duplications and divergence (Martin and Paz-Ares, 1997, Jin and Martin, 1999, Feller et al., 2011). Results from the EST analysis identified about 30 MYB genes in *Petunia hybrida* (Avila et al., 1993), over 80 in *Zea mays* (Rabinowicz et al., 1999) and 200 in the genus *Gossypium* (Cedroni et al., 2003). The systematic analysis of the genomes of *Arabidopsis thaliana* and *Oryza sativa* detected 198 and 183 in each (Yanhui et al., 2006).

Many studies have been performed in the plant MYB proteins and now the understanding of their roles are, though still not very clear, much better. Their involvement was found in many processes. For example, the cell division cycle 5 (CDC5), a cell cycle regulator, is critical for G2-M transition in *A. thaliana* just as it is in yeast and animals (Lin et al., 2007). It was first found in *Schizosaccharomyces pombe* and confirmed to be conserved in fungi, animals, and plants (Ohi et al., 1994, Ohi et al., 1998). Later, it was hypothesized to form a complex involved in the process of innate immunity which seems to be conserved across plant and animal kingdoms (Palma et al., 2007). Although the mechanisms of the AtCDC5 involvement in growth and immunity are still unclear, it was suggested that CDC5 might function as a transcription factor of the microRNAs, or act in the posttranscriptional processing of the primary miRNAs. The pleiotropic effect of AtCDC5 fits the feature of microRNAs involvement in different biological processes (Zhang et al., 2013).

As shown in the case of CDC5, cell cycle is one of the many biological processes that are regulated by MYB proteins (Weston, 1998, Oh and Reddy, 1999). MYB11 is another MYB protein that is crucial for cell cycle progression (Petroni et al., 2008), and these are just two of many. Many studies have shown the critical rules of some MYB proteins in animals, like the B-myb (Lam and Watson, 1993, Joaquin and Watson, 2003) and C-myb (Nakata et al., 2007). Their control over G2-M cell cycle transition is essential and studies in tobacco cells seem to suggest it is also the case in plants (Ito et al., 1998, Ito et al., 2001, Araki et al., 2004). The G2-M phase-specific B-type cyclin genes (*CYCB1*) have a mitosis-specific activator (MSA) element in their promoters (Ito et al., 1998). It is the target of three 3R MYB proteins in tobacco, NtMYBA1, NtMYBA2, and NtMYBB. The first two genes activated the MSA-containing promoters in the transient assays, while the activation was reduced at the presence of NtMYBB (Ito et al., 2001). Their homologs in *A. thaliana*, MYB3R1 and MYB3R4, which positively regulate cytokinesis, seem to

activate the G2-M phase-specific genes like B2-type cyclin (*CYCB2*), *CDC20.1*, and *KNOLLE* (*KN*). Their promoters contain MSA-like motifs (Haga et al., 2007). Although there is no evidence of their interaction with AtCYCB1;1, the rice homolog OsMYB3R2 was proven to bind to OsCYCB1;1, an MSA containing G2-M phase-specific gene, during chilling stress (Ma et al., 2009).

And sometimes they seem to work in association. For example, in *A. thaliana*, a member of the MYB proteins, GLABROUS1 (GL1) is believed to be in control of the cellular differentiation of trichomes (Oppenheimer et al., 1991, Payne et al., 1999, Payne et al., 2000). Its paralogous gene, MYB23, has shown a partially redundant role in controlling trichome morphogenesis and initiation (Kirik et al., 2005). Another case is MYB33 and MYB65, they redundantly facilitate anther development although the double mutant sterility is conditional (Millar and Gubler, 2005). Their specific expression in developing anthers is regulated by miR159 (Allen et al., 2007). MALE STERILE 1 (MS1), which targets MYB99, is known to regulate tapetal and pollen development and affect fertility (Wilson et al., 2001, Alves-Ferreira et al., 2007, Ito et al., 2007, Yang et al., 2007a). Other important MYB members related to anther or pollen development are MYB80/103 (Li et al., 1999, Higginson et al., 2003, Zhu et al., 2010, Phan et al., 2011, Phan et al., 2012, Xu et al., 2014b), MYB26 (Steiner-Lange et al., 2003, Yang et al., 2007b, Nelson et al., 2012), and TDF1 (Zhu et al., 2008, Gu et al., 2014). Not only male fertility, female fertility is also controlled by the MYB family. MYB98 is responsible for pollen tube guidance and synergid cell differentiation (Kasahara et al., 2005).

Apart from what has been discussed above, MYB genes also regulate a lot of other biological processes. Sometimes one gene regulates multiple pathways, and sometimes multiple genes control one feature. In *A. thaliana*, *MYB61* is required for germination and seedling establishment (Penfield et al., 2001), meanwhile it is also related to the ectopic lignification and dark-photomorphogenesis (Newman et al., 2004). Another R2R3 MYB gene, *LAF1*, is involved in photomorphogenesis as well (Ballesteros et al., 2001, Seo et al., 2003). *CAPRICE* (*CPC*) and *WEREWOLF* (*WER*) act in opposition to each other, determining the root epidermal cell differentiation together (Wada, 1997, Lee and Schiefelbein, 1999, Wada, 2002). *MYB4* is responsive to UV stress (Hemm et al., 2001) and the expression of *MYB102* is linked to wounding and osmotic stress (Denekamp, 2003). Studies in other species also revealed many MYB gene regulations. *NtMYB2* is activated by wounding and elicitors (Sugimoto et al., 2000). *ROUGH*

SHEATH 2 (*RS2*) in *Z. mays* and *Blind* in *Solanum lycopersicum* both control the formation of meristem development (Timmermans, 1999, Schmitz et al., 2002). There is even the case in *O. sativa* of the rice telomere-binding protein 1 (RTBP1), a MYB protein that binds to the double-stranded telomeric DNA (Yu et al., 2000).

1.4 The GAMYB-like family

The R2R3 MYB proteins compose the predominant MYB family in plants (Jin and Martin, 1999). 126 members were found in *Arabidopsis thaliana* alone (Yanhui et al., 2006). A lot of them seem to act as transcription factors that are particularly important to plants (Martin and Paz-Ares, 1997, Romero et al., 1998). In addition, they are only found in plants (Riechmann, 2000), and are considered to have evolved from an R1R2R3 MYB protein ancestor by losing the first repeat (Braun and Grotewold, 1999). Therefore, their regulation of plant-specific processes suggests that the R2R3 MYB family has played an important role in plant evolution (Stracke et al., 2001). Also, the low redundancy of the family members shows the evolution of transcriptional regulation on different temporal and spatial expression in plant developmental processes (Stracke et al., 2001).

A MYB gene was found to be involved in gibberellin (GA)-regulated gene expression in barley. This *HvGAMYB* is a transcription factor which regulates a hormone signalling pathway (Gubler et al., 1995). In *A. thaliana*, seven genes were considered to be homologs of *HvGAMYB*, including *MYB33*, *MYB65*, *MYB81*, *MYB97*, *MYB101*, *MYB104*, and *MYB120* (Jiang et al., 2004, Dubos et al., 2010). Although the previously discussed phylogenetic study showed that *DUO1* is closely related to this R2R3 MYB family also known as the *GAMYB* family (Jiang et al., 2004), further research suggested that *DUO1* may form a clade which is distinct from the *GAMYB* clade that includes those seven *A. thaliana* members (Dubos et al., 2010). All eight genes bear a feature that is common among R2R3 MYB proteins, the substitution of one tryptophan at the beginning of the R3 repeat (Romero et al., 1998). The roles of the GAMYB family proteins are intriguing, as apart from MYB81 and MYB104 which were not previously studied, the remainders all have functions related to male reproduction (Millar and Gubler, 2005, Allen et al., 2007, Liang et al., 2013). Further evidence that links the GAMYB family to reproduction is the fact that the GA pathway has a strong impact on fertility in *A. thaliana* and *Oryza*

sativa by affecting the microspore development (Cheng et al., 2004, Plackett et al., 2011, Plackett et al., 2012).

The seven members of the GAMYB family are all expressed in the male gametophyte (Dubos et al., 2010, Liang et al., 2013). MYB33 and MYB65 are involved in anther development (Millar and Gubler, 2005, Allen et al., 2007), while MYB97, MYB101, and MYB120 are responsible for pollen tube-synergid interaction as male factors (Liang et al., 2013). MYB97, MYB101, and MYB120 are also expressed exclusively in mature pollen and localized in the nucleus. They form one branch with 32 % amino acid sequence identity (Leydon et al., 2013, Liang et al., 2013, An et al., 2014). The unpublished data from Dr Borg shows that the MYB domain of DUO1 alone is able to locate in the nucleus, this suggests the closely related GAMYB members may all have the same nucleus-locating property. MYB101 has the highest expression level among the three. MYB97 and MYB101 may function as transcription activators, but MYB120 does not (Renak et al., 2012, Liang et al., 2013, An et al., 2014). MYB101, along with MYB33, is a regulator of ABA signalling (Reyes and Chua, 2007, Kim et al., 2008b, Daszkowska-Golec et al., 2013). The whole family is regulated by miR159 like DUO1 (Jiang et al., 2004, Allen et al., 2007, Brownfield et al., 2009a, Allen et al., 2010).

The *A. thaliana* triple mutant line (*myb97-1 myb101-1 myb120-3*) has a highly reduced fertility. The expression of MYB97, MYB101, and MYB120 with the promoter of *MYB101*, or even promMYB101:MYB33 and promMYB101:MYB81 all restored the fertility completely. The whole family are functionally redundant in pollen tube reception (Liang et al., 2013).

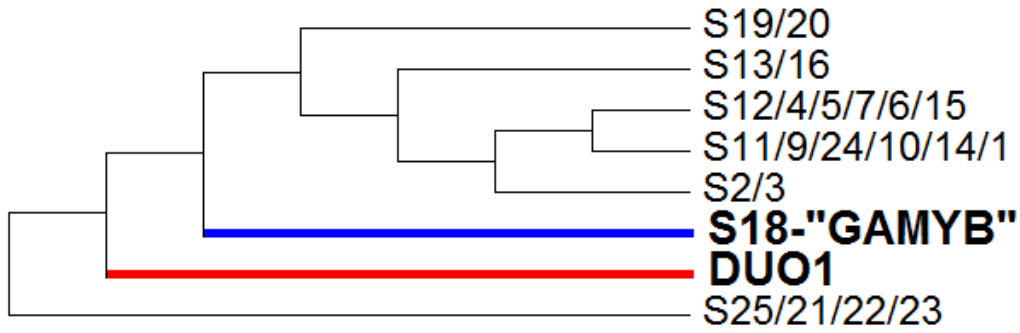


Figure 1. 5 Phylogeny of the R2R3 MYB proteins in *A. thaliana*.
DUO1 is closely related to the GAMYB family according to the phylogenetic study of the R2R3 MYB proteins (Dubos et al., 2010). Picture adapted from the same research.

1.5 The regulation of microRNA

The history of the microRNA study is relatively short. It began with the discovery of the fact that *lin-4*, a gene that regulates the development timing in *Caenorhabditis elegans*, produces a pair of short noncoding RNAs instead of coding for a protein (Lee et al., 1993). One of them is 61 nt in length, and seems to be a precursor of the other small RNA, which is about 22 nt. Their complementarity towards the repression region of another gene, *lin-14*, pushed forward the idea of a novel regulatory method (Wightman et al., 1991, Wightman et al., 1993, Lee et al., 1993).

Later, the 22 nt *lin-4* RNA, along with a bunch of other similar tiny regulatory RNAs, were classified as microRNAs, or miRNAs (Lagos-Quintana et al., 2001, Lau et al., 2001, Lee and Ambros, 2001). A huge number of microRNAs have since been found in animals, plants, and even viruses (Griffiths-Jones et al., 2008). Hundreds of miRNAs were indentified in *C. elegans* (Ruby et al., 2006), *Drosophila melanogaster* (Ruby et al., 2007), *Homo sapiens* (Landgraf et al., 2007), and *Arabidopsis thaliana* (Adai et al., 2005). Computational predictions indicate miRNAs control an immensely wide range of biological processes across kingdoms (Rhoades et al., 2002, Enright et al., 2003, Lewis et al., 2003, Stark et al., 2003, John et al., 2004, Kiriakidou et al., 2004, Rajewsky, 2006, Alves et al., 2009). Some of the confirmed miRNAs functions and their targets include the regulation of cell proliferation (Brennecke et al., 2003), cell death and metabolism (Xu et al., 2003), cell cycle (Vasudevan et al., 2008), cell differentiation (Chen, 2004), neuronal development (Johnston and Hobert, 2003), meristem development (Emery et al., 2003), and flower development (Aukerman and Sakai, 2003, Chen, 2004). However, there is still much to know about these huge regulatory networks (Bartel, 2004, He and Hannon, 2004).

In plants, most miRNAs are not so different from their animal counterpart, which indicates the conserved nature of the miRNAs (Voinnet, 2009). However, most plant miRNA genes are intergenic, instead of within introns or exons like in animals (Kim, 2005, Zhang et al., 2008). Meanwhile, some miRNA families seem to be conserved across all land plants, including miR156, miR160, miR319, and miR390. These miRNA families regulate transcription factors that control multiple biological processes, some of which are among the MYB protein family (Garcia, 2008). Interestingly, miRNAs in the unicellular alga, *Chlamydomonas reinhardtii*, seem to have evolved with the ones in

multicellular plants and animals separately (Molnar et al., 2007). Furthermore, there are 21 families conserved among angiosperms but are not present in the moss, *Physcomitrella patens* (Axtell and Bowman, 2008). Notably, although evidence indicates that miR159 and miR319 are closely related (Palatnik et al., 2007, Li et al., 2011b), there is no report to suggest that the former one is present in moss like the latter one. Another difference between the miRNAs from plants and animals is the complementarity-cleavage relationship. While this relationship is rather complicated in animals thus remains a challenge for target prediction (Yekta et al., 2004, Davis et al., 2005), many miRNA targets in plants can simply be predicted by checking the extensive complementarity (Rhoades et al., 2002). There are many websites that can provide a prediction with confidence (Voinnet, 2009, Bonnet et al., 2010).

1.6 The miR159 regulation of the GAMYB family

In 2003, a mutant type of *Arabidopsis thaliana* with pleiotropic developmental defects was analysed. Three isoforms of miR159 (a, b, and c) were predicted to target the MYB genes like *MYB33*, *MYB65*, and *MYB104* (Palatnik et al., 2003). Later, the miR159 expression level was proven to be modulated by the gibberellic acid (GA) during anther development (Achard et al., 2004) and abscisic acid (ABA) during seed germination (Reyes and Chua, 2007). It suggested that miR159 might function as a regulator of the GAMYB proteins (Achard et al., 2004), and then was proven to mediate the cleavage of *MYB33* and *MYB101*, two members of the *GAMYB* family, *in vitro* and *in vivo* (Reyes and Chua, 2007). Then, studies showed that miR159 regulates the *GAMYB* transcript levels in vegetative tissues, completely silencing *MYB33* and *MYB65* (Alonso-Peral et al., 2010, Alonso-Peral et al., 2012). Further research has also shown the miR159-guided cleavage on *MYB81*, *MYB120*, and *DUO1* (*MYB125*) (Allen et al., 2010). Although it was also pointed out that miR159 regulation *in vivo* is limited to only *MYB33* and *MYB65* (Allen et al., 2007, Allen et al., 2010), an independent study on *MYB120*, which is mainly transcribed in anther and pollen but negligibly in other tissues, indicates otherwise (Winter et al., 2007, Li and Millar, 2013). Nevertheless, with the evidence of miRNA transcriptional regulations within the male germline (Grant-Downton et al., 2013), it is totally possible for miR159 to act as a safety measure for the other genes in case of the leaky transcriptions, a situation often seen in animals (Allen et al., 2010).

Considering the conserved nature of both the MYB genes and the miRNAs, it was not surprising that miR159 in another species, *Sinningia speciosa*, also regulates its target *SsGAMYB*, which controls flowering time (Li et al., 2013b). However, it is rather interesting that the miR159 in tomato, *Solanum lycopersicum*, targets a non-MYB-related gene, *SGN-U567133*, which is involved in leaf and flower development (Buxdorf et al., 2010). This leads to a more important question, the regulation mechanism of miR159. Functional specializations of the miR159 and its close family miR319 in *A. thaliana* are through different mechanisms. The expression level of miR319 restricts its influence on MYB mRNAs, while the sequence of miR159 blocks its interaction with *TCP* (Palatnik et al., 2007). Evidence shows that even though complementarity is the number one factor of the regulating effect, mismatch is allowed to have an efficient outcome (Li et al., 2014b). This means a microRNA can target many slightly different sequences. In addition, miR159 silences its target through both cleavage and non-cleavage mechanism (Li et al., 2011b, Li et al., 2014b).

1.7 Sequence-specific DNA binding transcription factors

The process of transcription *per se* from DNA to messenger RNA (mRNA) is performed by RNA polymerase (Cramer et al., 2001). A transcription factor (TF) is a DNA binding protein that, on its own or with other proteins, controls this process (Karin, 1990, Latchman, 1997), acting as an activator (Buratowski et al., 1989, Conaway and Conaway, 1993, Roeder, 1996, Nikolov and Burley, 1997, Gill, 2001), a repressor(Lee and Young, 2000), or both at the same time (Adkins et al., 2006, Ikeda et al., 2009). Its regulation sequence specificity is determined by one or more DNA recognition domains known as DNA-binding domains (DBDs) (Mitchell and Tjian, 1989, Ptashne and Gann, 1997), a characteristic that separate a transcription factor from other gene regulators (Brivanlou and Darnell, 2002).

There are many transcription factor families, including but not limited to basic helix-loop-helix (bHLH) (Massari and Murre, 2000), basic-leucine zipper (bZIP) (Vinson et al., 1989), helix-turn-helix (HTH) (Brennan and Matthews, 1989), homeodomain proteins (Gehring, 1992), zinc fingers (Klug and Rhodes, 1987, Laity et al., 2001), Cys2His2 zinc fingers (Wolfe et al., 2000, Klug, 2010), and MYB (myeloblastosis) proto-oncogene proteins (Saikumar et al., 1990). Transcription factor networks are found

in all living organisms because of their importance in gene regulations (Lemon and Tjian, 2000, Riechmann, 2000, Huffman and Brennan, 2002, Lee et al., 2002, Babu et al., 2004). The number of transcription factors also seems in positive correlation to the genome size of an organism (van Nimwegen, 2003), which agrees with the idea of certain gene balance hypotheses (Edger and Pires, 2009).

1.8 Nine-amino-acid transactivation domain family

A trans-activation domain (TAD) contains a binding site that recruits other transcriptional regulators (Warnmark et al., 2003). One common chemical feature shared by most TADs is the amino acids hydrophobicity (Drysdale et al., 1995, Sullivan et al., 1998), acidity (Hope and Struhl, 1986, Hope et al., 1988, Gill and Ptashne, 1987, Sadowski et al., 1988), or a combination of both (Ma and Ptashne, 1987, Regier et al., 1993, Triezenberg, 1995, Sainz et al., 1997).

A large group of those TADs is the nine-amino-acid transactivation domain (9aaTAD), which can be seen in a superfamily of eukaryotic transcription factors including Gal4, Gcn4, p53, and VP16 (Piskacek et al., 2007). The 9aaTADs of those most well studied proteins have shown to interact with other general coactivators like p300 and TAF (Gu and Roeder, 1997, Uesugi et al., 1997). The induced complex will then activate its target (Liu et al., 1999).

While the most studied 9aaTAD protein is no doubt p53, VP16 however is one of the most used tool as a trans-activator in plant molecular biology (Zuo et al., 2000, Lohmann et al., 2001, Storgaard et al., 2002, Silveira et al., 2007, Ikeda et al., 2009, Hanano and Goto, 2011, Aguilar et al., 2014). VP16, also known as herpes simplex virus protein vmw65, is a key activator for HSV lytic infection (Herrera and Triezenberg, 2004). While the core structure of the protein is in charge of DNA recognition, the C-terminal 9aaTAD recruits Oct-1 and HCF-1 (Liu et al., 1999, Wysocka and Herr, 2003).

1.9 Bioinformatics and phylogenetics

The history of phylogenetic trees goes far beyond the discovery of DNA. Darwin's notebook from 1837 has an evolutionary tree of life. However, traditionally most of the

phylogenetic trees are drawn based on systematics or taxonomy. With the help of DNA sequencing technology, phylogenetics is able to study the evolutionary relationships among species or populations with molecular data (Slowinski and Page, 1999).

Today, the application of phylogenies has extended outside the tree of life. They are applied in the study of the relationships between gene or protein homologs (Maser et al., 2001), cell lineages (Salipante and Horwitz, 2006), evolution of pathogens (Marra et al., 2003, Grenfell et al., 2004), population history (Edwards, 2009), and even the evolution of languages (Gray et al., 2009). Lately, molecular phylogenetics has been used in the identification of genes (Kellis et al., 2003), miRNAs (Pedersen et al., 2006), regulation factors (Lindblad-Toh et al., 2011), metagenomics (Brady and Salzberg, 2011), ancestor genomes reconstruction (Paten et al., 2008, Ma, 2011), and genome interpretation (Green et al., 2010, Gronau et al., 2011, Li and Durbin, 2011).

After constructing a phylogenetic tree of a certain gene family, one thing that can be done is to detect the positive selection on certain sites. Positive selections can be identified through the comparison of the synonymous and non-synonymous substitution rates (Nei and Kumar, 2000). There are many methods that exist to test the selection for the gene sequences on certain branches (Yu and Irwin, 1996, Messier and Stewart, 1997, Zhang et al., 1997, Yang, 1998, Zhang et al., 1998), certain codon sites on the whole tree (Nielsen and Yang, 1998, Suzuki and Gojobori, 1999, Yang et al., 2000), or certain codon sites on certain branches (Yang and Nielsen, 2002). Although technical problems still present a challenge in errors like false detections (Zhang, 2004), revised new models are being created to improve the algorithms as well, giving a promising future in the field (Zhang et al., 2005).

Another part of bioinformatics in which has been put a lot of efforts is the estimation of the evolutionary timescale. Traditionally, it was established solely through palaeontologists comparing fossil record (Clarke et al., 2011). The utilization of the molecular clock has changed this situation. The method has been developing rapidly in the last twenty years (Takezaki et al., 1995, Sanderson, 1997, Thorne et al., 1998, Rambaut, 2000, Sanderson, 2002, Drummond et al., 2006), it can even be done in the absence of a molecular clock (Sanderson, 2003). However, the result can be influenced by many factors, as the calculation is completely based on the DNA polymorphism,

which is susceptible to conditions like positive selections (Eyre-Walker and Keightley, 2009, Keightley and Eyre-Walker, 2012).

1.10 The origin and phylogeny of land plants

The earliest surviving land plants, or Embryophyta, emerged around 480 million years ago (mya) in the Ordovician period, and later diversified and shaped the world as we know it today (Kenrick and Crane, 1997, Kenrick et al., 2012, Wellman et al., 2003, Steemans et al., 2009, Rubinstein et al., 2010, Magallon et al., 2013). The timeline of early land plants aligned with the ending of the "Snowball Earth" period (ca. 650 mya), akin to the famous "Cambrian explosion" in the animal kingdom.

Our insight into plant phylogeny today no longer depends on morphological studies, but is largely contributed by bioinformatic approaches (Slowinski and Page, 1999). The current phylogenetic tree of green plants is developed through the comparison of plasmid genes (Bremer et al., 2003, Bremer et al., 2009, Jansen et al., 2007, Moore et al., 2007, Moore et al., 2010, Ruhfel et al., 2014), mitochondrial genes (Qiu et al., 2010), ribosomal genes (Burleigh et al., 2009, Soltis et al., 2011), and nuclear genes (Timme and Delwiche, 2011, Burleigh et al., 2011, Wickett et al., 2014). While these studies agree with the early taxonomy hypothesis in general with a few changes (Cronquist, 1988, Thorne and Reveal, 2007), some questions are still left to be answered.

The most significant one is the origin of land plants. Although Charophyta and Embryophyta (collectively known as the Streptophyta) are proven to be monophyletic (Surek et al., 1994, Kenrick and Crane, 1997, Lemieux et al., 2007, Qiu and Lee, 2000, Qiu et al., 2006), the algal lineage branches hierarchy in relation to the land plants remains unclear (Turmel et al., 2006, Wodniok et al., 2011, Laurin-Lemay et al., 2012, Timme et al., 2012).

Physiological and morphological comparisons suggest that the sister group of Embryophytes is among Charales, Coleochaetales, and some members of the Zygnematophyceae (Pickettheaps and Wetherbee, 1987, Galway and Hardham, 1991). A few studies supported this idea (Bhattacharya and Medlin, 1998, Graham et al., 2000), suggesting a clade including Charales and Embryophyta is sister to Coleochaetales (Karol et al., 2001, Lewis and McCourt, 2004). However, more recent phylogenomic

analyses have cast doubts upon this. While Coleochaetales has less support (Finet et al., 2012), there is a large amount of evidence demonstrating that Zygnematophyceae (Wodniok et al., 2011, Timme et al., 2012, Zhong et al., 2013, Ruhfel et al., 2014, Civan et al., 2014, Wickett et al., 2014, Davis et al., 2014), or a clade with both lineages (Turmel et al., 2006, Turmel et al., 2007, Chang and Graham, 2011, Leliaert et al., 2012), is the closest relative to all land plants.

The question of origin does not stop there. Even though we know as a fact that bryophytes, including mosses, liverworts, and hornworts, are the results of the earliest diversification events within embyrophytes (Qiu et al., 1998, Qiu et al., 2006, Qiu et al., 2007, Nickrent et al., 2000, Nishiyama et al., 2004, Shaw et al., 2011), their relationships with the tracheophytes and with each other remain unsolved.

Some believe that byrophytes are monophyletic (Nishiyama et al., 2004, Cox et al., 2014), some are convinced the three form a grade with hornworts sister to tracheophytes (Qiu et al., 1998, Qiu et al., 2006, Qiu et al., 2007, Groth-Malonek et al., 2005, Chang and Graham, 2011), while some think that only mosses and liverworts are monophyletic (Nickrent et al., 2000, Lemieux et al., 2007, Karol et al., 2010). There is little surprise that the topology of the three branches with tracheophytes varies among these studies.

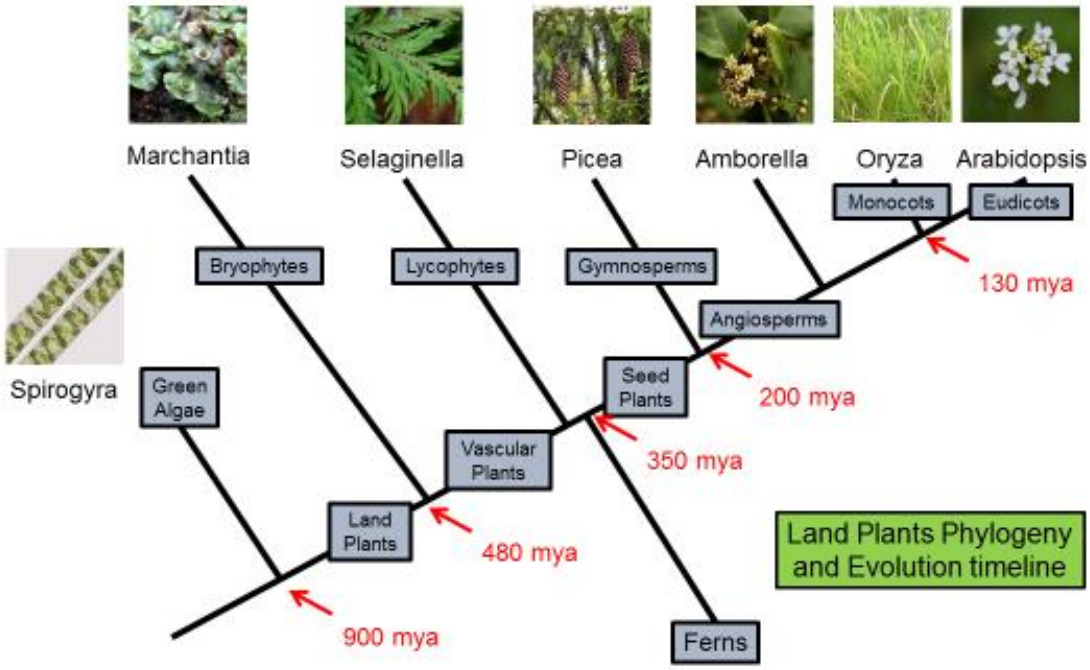


Figure 1. 6 All embryophytes diverged from the streptophyte algae (charophytes). The times on the nodes are the approximate dates based on literature. It is important to note that the enormous span of uncertainty regarding the divergence time. For example, although we know that the earliest living divergent land plant group (bryophytes) branched out around 450 to 480 million years ago (mya), they diverged from the streptophyte alga between 870 to 1042 mya (Clarke et al., 2011, Magallon et al., 2013). The first flowering plant evolved around 140 to 250 mya and no fossil older than 130 mya has been found so far (Clarke et al., 2011, Magallon et al., 2015, Foster et al., 2017, Sauquet et al., 2017). Picture credits see Appendix.

1.11 Bryophytes as model species

The establishment and study of model organisms have contributed greatly to our understanding of evolution today, including the phylogeny and timescale of life (Hedges, 2002). To solve the problems surrounding the origin of land plants, much effort has been put into the study of bryophytes (Nickrent et al., 2000, Hedges, 2002, Nishiyama et al., 2004, Qiu et al., 2006, Chang and Graham, 2011, Cox et al., 2014). The moss *Physcomitrella patens* and recently the liverwort *Marchantia polymorpha* have become the model organisms of bryophytes (Cove et al., 1997, Cove, 2005, Cove et al., 2006, Takenaka et al., 2000, Chiyoda et al., 2006, Chiyoda et al., 2008, Goffinet and Shaw, 2009).

The moss *P. patens* has been the subject of many molecular studies including hormone synthesis pathways (Lindner et al., 2014), light response including cell regeneration morphogenesis (Jenkins and Cove, 1983, Yamawaki et al., 2011, Ranjan et al., 2014), cell cycle and cell fate regulation (Tanahashi et al., 2005, Jang et al., 2011, Aoyama et

al., 2012), and plant evolution (Nishiyama et al., 2003, Rensing et al., 2008). Recently, there is even some attempt to move it beyond just the model of molecular biology (Müller et al., 2015). However, moss has the most attention on its unique feature in transformation (Schaefer et al., 1991, Kammerer and Cove, 1996, Wood et al., 2000), which has given people easy access to gene targeting (Schaefer and Zryd, 1997, Schaefer, 2002). Although the idea of gene targeting through homologous recombination (HR) has been around for a long time (Capecchi, 1989), it was impractical in most high plants due to their low efficiency (Hanin and Paszkowski, 2003). Before the development of the gene-editing technique CRISPR (Ran et al., 2013, Schiml et al., 2014, Hsu et al., 2014, Bortesi and Fischer, 2015), moss allowed researchers to utilize the benefit of HR-mediated genetic engineering (Schaefer and Zryd, 1997, Schaefer, 2002, Trouiller et al., 2007).

A lot of genes involved in the moss homologous recombination process have been identified (Trouiller et al., 2006, Kamisugi et al., 2012), among them the two homologs of *RAD51* (Ayora et al., 2002). *RAD51* regulates homologous recombination and HR-mediated repair, its homologs have been identified in yeast, vertebrates, and plants, all of which play a conserved role (Sung and Robberson, 1995, Krogh and Symington, 2004, Baumann and West, 1998, Ayora et al., 2002, Holthausen et al., 2010, Charlot et al., 2014). However, sequence analysis put the two *P. patens RAD51* genes in a clade separated from other plants, and the difference from other multicellular eukaryotes is possibly responsible for the high efficiency of gene targeting in moss (Markmann-Mulisch et al., 2002). This hypothesis is further backed up by independent knockout experiments, which demonstrate that both genes are contributing in the process of homologous recombination, maintenance of genome integrity, and resistance to DNA damaging, with one gene being dominant (Ayora et al., 2002, Markmann-Mulisch et al., 2007, Schaefer et al., 2010, Charlot et al., 2014). However, it remains unclear whether there is a fundamental difference between the homologous recombination mechanisms of moss and other plants.

The study of the liverwort *M. polymorpha* can be dated back to as early as the ancient Greeks, but regrettably little literature is available (Bowman, 2015). Its haploidy and the separation of genders present a great opportunity for molecular genetic study, and its transformation system has been developed over the last two decades (Takenaka et al., 2000, Chiyoda et al., 2006, Chiyoda et al., 2008, Ishizaki et al., 2008). Gene targeting approaches have also been used in some molecular studies (Ishizaki et al., 2013, Ueda et

al., 2014). Although previously only the chloroplast and mitochondrial data were publicly available (Ohyama et al., 1986, Oda et al., 1992), the *M. polymorpha* genome has been sequenced (Bowman et al., 2017). With new techniques like CRISPR, *M. polymorpha* has already become another key model species in bryophytes.

1.12 Angiosperms and whole genome duplications

The sudden domination of the angiosperms was coined "an abominable mystery" by Charles Darwin in 1879. Now the most abundant branch of land plants with over 300,000 species, the rise of angiosperms is around the Cretaceous-Tertiary extinction event (Schneider et al., 2004, Renne et al., 2013), partially thanks to the advantages associated with polyploidy which accelerated the creation of new genes (De Bodt et al., 2005, Soltis et al., 2008, Soltis et al., 2009, Edger and Pires, 2009, Fawcett et al., 2009). Being one of the most powerful forces in evolution, whole-genome duplication (WGD), or polyploidy events influenced the history of fungi (Kellis et al., 2004), animals (Blomme et al., 2006, Kassahn et al., 2009), and plants (Edger and Pires, 2009, Muhlhausen and Kollmar, 2013).

Most angiosperm linages show evidence of several WGD events (Jiao et al., 2011). In core eudicots, while there is very likely a common triplication event (known as γ), the study on *Arabidopsis thaliana* shows two more possible duplication events (α and β) within the Brassicaceae family (Vision et al., 2000, Blanc et al., 2003, Bowers et al., 2003, Jaillon et al., 2007, Lyons et al., 2008, Tang et al., 2008a, Tang et al., 2008b, Barker et al., 2009). In monocots, two such duplication events (ρ and σ) have been demonstrated (Paterson et al., 2009, Tang et al., 2010). There is also evidence suggesting an even earlier duplication event (ε) probably shared by all angiosperms (Vision et al., 2000, De Bodt et al., 2005, Cui et al., 2006, Soltis et al., 2008, Soltis et al., 2009), which is further supported by the sequencing of the *Amborella trichopoda* genome (Albert et al., 2013).

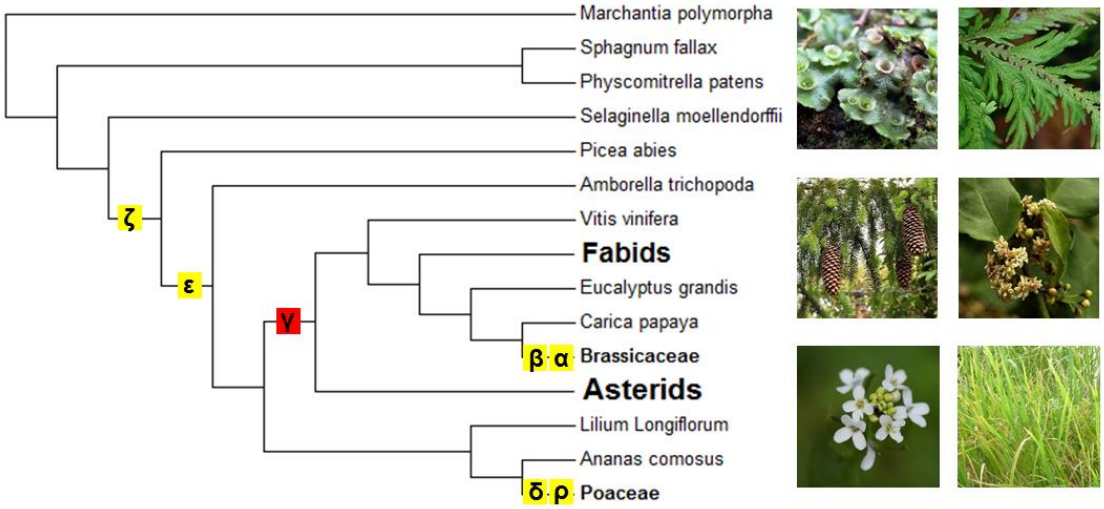


Figure 1. 7 Ancestral polyploidy events in seed plants and angiosperms.
The yellow boxes indicate known whole genome duplication events and the red box indicates a whole genome triplication event. Angiosperms, especially for members of the family Brassicaceae and Poaceae, have experienced several polyploidy events. Diagram adapted from Jiao et al., 2011.

Regarding the fate of nuclear genes following duplication events, there are many hypotheses including the Gene Balance Hypothesis (or Dosage Balance Hypothesis), Gain-of-function hypothesis, Subfunctionalization, Increased Gene Dosage Hypothesis, and Functional Buffering Model (Edger and Pires, 2009). A regulator gene connected with a complex network, such as a transcription factor, is sensitive to the imbalance in the concentration. A change in one such unit could lead to a catastrophic effect or at least a decreased fitness. This is known as the Gene Dosage Hypothesis (Veitia, 2002, Veitia, 2005, Veitia et al., 2008). This hypothesis points out that for an important transcription factor regulating a network, retention of another copy after duplication is very unlikely unless all genes in the whole network increase their dosages simultaneously.

1.13 Aims and Objectives

Recent efforts have identified a regulatory network controlling the pollen mitosis II process in the flowering plants. As described in Section 1.2, the centrepiece *DUO POLLEN 1* (*DUO1*) is required for both the G2-M cell cycle transition and the sperm cell differentiation. Much more information has been gathered for the *DUO1* network in the model plant *Arabidopsis thaliana*. However, although it was suspected that almost all flowering plant species should have at least one copy of this gene, only a handful of *DUO1* homologs have been identified previously. Neither was any functional

information available for any of these homologs. The scope of the *DUO1* conservation was not fully investigated or understood.

The aim of this thesis was to understand the sequence and functional conservation of *DUO1*. The pursuit that would lead to the understanding of the origin of *DUO1* and its evolution in the plant lineages.

The first major objective, which is discussed in Chapter 3, was to determine the sequence conservation of the DUO1 homologs using available sources. These data also provided an opportunity to look into the protein secondary and tertiary structure conservation among them. The expression patterns of these homologs suggested that DUO1 functions have been related to the male germline development since bryophytes. The large collection of DUO1 sequences has revealed its algal origin in the Charophyta and its sister group, the GAMYB family proteins. The two male reproduction related MYB clades have very different fates regarding the subsequent proliferation following the divergence.

A second major objective in Chapter 4 was to test the functional conservation of the DUO1 homologs using experimental methods. These typically included *in planta* complementation and dual-luciferase reporter assays. It was demonstrated that different DUO1 homologs from angiosperms could largely substitute one another functionally and they are structurally highly similar in both the MYB domain and the C-terminal end. In contrast, although DUO1 homologs from pre-angiosperms have a highly similar MYB domain that binds to the same target DNA sequence, the C-terminal end is visibly different from the angiosperm DUO1 sequences. They were unable to replace the native DUO1 functions and promote sperm differentiation in *A. thaliana*, either. The emergence of the angiosperm was a watershed for the functional changes of DUO1.

The next chapter focus on the impact of the miR159 regulation on *DUO1*. The miR159 binding site encodes a signature supernumerary residue for the DUO1 protein MYB domain. The finding suggested that *DUO1* was not under the regulation of the microRNA until the core angiosperm group. The regulation of the *GAMYB* family on the other hand predated the bryophyte, before the separation of miR159 and miR319 clades. Therefore, it is clear that the miR159 binding sites for *DUO1* and *GAMYB* clades developed independently.

The objective of Chapter 6 was to look at the differences of the C-terminal regions between the angiosperm and pre-angiosperm DUO1 homologs. The motifs at the DUO1 C-terminal region of the angiosperms were found crucial for the high level of the target promoter activation. Strangely, changing this region in the pre-angiosperm DUO1 protein examined failed to yield any positive result. It is proposed that the variable region linking the MYB DNA-binding domain and the C-terminal activation motif plays an important role despite the lack of any apparent conservation. Nevertheless, the DUO1 C-terminal region showed some evidence of being a member of the famous nine-amino-acid transactivation domain.

Collectively, the findings in this thesis provided an understanding of the conservation of a key transcription factor in the plant male germline gene network, at an evolution level and a molecular level.

Chapter 2 Materials & Methods

2.1 Purchase of materials

Chemicals and materials were ordered from Melford Laboratories, Sigma-Aldrich, Thermo Fisher Scientific, and Promega. Commercial kits were ordered from QIAGEN, Sigma-Aldrich, and Omega Bio-tek. Enzymes and reagents were ordered from Invitrogen, Sigma-Aldrich, New England Biolabs (NEB), and Bioline.

2.2 Bacteria culture

Bacteria strains and the concentration of antibiotics used for selections are listed.

Working Concentration (µg/µl)		
	E. coli	*A. tumefaciens*
Antibiotics/Strains	α-select; DB3.1	GV3101
Ampicillin (AMP)	50	50
Chloramphenicol (CM)	25	25
Gentamicin (GENT)	N/A	50
Kanamycin (KAN)	50	50
Rifampicin (RIF)	N/A	25
Spectinomycin (SPEC)	100	100

Escherichia coli (*E. coli*) and *Agrobacterium tumefaciens* (*A. tumefaciens*) were cultured in Luria Bertani Broth medium (LB Broth). To grow the bacteria on plates, 1.5 % (w/v) bacto-agar was added to it before autoclaving. The pH was adjusted to 7.2 with 1 M NaOH.

2.3 DNA amplification

DNA amplifications were performed using the Polymerase Chain Reaction (PCR). The PCR conditions varied according to the size of the product and T_m of the primers. Double stranded DNA was denatured at either 96 °C with BioTaq (Bioline)/KAPA Taq (Sigma-Aldrich) or 98 °C with Velocity (Bioline)/Phusion (NEB). The annealing temperature T_a was set 5 – 10 °C below the T_m. Extension was set at 68 – 72 °C for 30 seconds per 1 kb

with BioTaq and 15 seconds per 1 kb with Velocity. The cycles were between 30 and 40. The thermocyclers used for the PCR reactions were the TProfessional Basic Gradient/Trio Thermocycler (Biometra).

PCR reactions for general purpose were performed with BioTaq DNA polymerase (Bioline). The relevant buffers used to make master mixes were supplied with the enzyme. The condition parameters for BioTaq reactions were: 96 °C for 2 minutes; 30 – 40 cycles of 96 °C for 30 seconds, 55 – 65 °C for 30 seconds and 68 – 72 °C for 30 seconds per 1 kb; and finally 68 – 72 °C for 5 minutes. For high fidelity PCR reactions, Velocity DNA polymerase (Bioline) was used. The denaturing process was changed from 96 °C to 98 °C, 30 seconds to 20 seconds; and extension time was decreased from 30 seconds per 1 kb to15 seconds per 1 kb. Other conditions remain the same.

2.4 DNA extraction and purification

Isolation of the plasmid DNA of *E. coli* cells was carried out using GenElute™ Plasmid Miniprep Kits (Sigma-Aldrich), E.Z.N.A® Plasimid Mini Kit, and E.Z.N.A® Plasimid Maxi Kit (Omega Bio-tek). PCR products were purified using either GenElute™ *PCR Clean-Up Kits* or GenElute™ Gel Extraction Kits (Sigma-Aldrich), E.Z.N.A® Cycle Pure Kit or E.Z.N.A® Gel Extraction Kit (Omega Bio-tek). Extraction of the gDNA from plant tissues was done using DNeasy Plant Mini Kit (QIAGEN).

2.5 Gateway® and other cloning methods

Unless specified, constructs were generated by MultiSite Gateway Technology (Invitrogen, UK) as described in Borg *et al*., 2011, using Gateway® recombination technology. This includes two steps: BP reaction to generate entry clones using donor vectors and purified PCR products, and LR reaction to generate target destination vectors using entry clones. Promoter entry clones were made in pDONRP4P1R, cDNAs in pDONR221, and fluorescent tags in pDONRP2R-P3. For example, the constructs pB7m34GW_promDUO1:AtDUO1-mCherry was generated using entry clones for the *DUO1* promoter, *AtDUO1* cDNA, and *mCherry*. The entry clones (10 µM each) were then used to build gene constructs by 3-part recombination into the T-DNA destination vector pB7m34GW (20 µM) (Karimi et al., 2002).

Otherwise, DNAs were amplified with restriction sites at both ends and then digested using the NEB high fidelity enzymes and CutSmart® buffer, before ligation with T4 DNA ligase.

For constructs used in the mammalian two-hybrid assays, plasmids were generated in the Protein Expression Laboratory (Protex) service from University of Leicester using vectors available through the service. The effectors were cloned into pLEICS-12 that contains the human cytomegalovirus (CMV) promoter and the reporter was cloned into pLEICS-13 with the *HTR10* promoter driving the firefly luciferase.

2.6 Gel electrophoresis

Identification of DNA fragments acquired through different processes was carried out by agarose gel electrophoresis. Agarose gels were made in 1 x TAE (40 mM Tris base, 20 mM glacial acetic acid, 1 mM EDTA) supplemented with 0.2 µg/ml ethidium bromide. Depending on the size of the DNA fragments, the concentration of agarose gels varied between 0.8 to 3 % (w/v). Samples were loaded into the gel wells after mixing with 3 x Orange G DNA loading buffer (3 % glycerol, 200 µg/ml Orange G) in a volume ranging between 5 - 50 µl depending on the application. The loaded gel was placed in an electrophoresis tank and a voltage from 100 - 150 V was applied for 30 - 60 minutes. The DNA fragments were then visualised under a UV transilluminator (BioDoc-ItTM System, UVP). The size and quantity were determined by comparing with the standard DNA ladders (New England Biolabs).

2.7 Plant materials and growth conditions

Arabidopsis thaliana plants were grown on soil in greenhouse conditions (21 to 25 °C) with a 16 h photoperiod or in growth chambers at 24 °C under continuous illumination (120 to 140 mmol/m^2/s with 60 % humidity). The *duo1-1* mutant line has been described previously (Durbarry *et al.*, 2005). The transgenic Nossen-0 *duo1-1*$^{+/-}$ *pK7m34GW_promHTR10:H2B-GFP*$^{+/+}$ marker line has been described previously (Brownfield et al., 2009a), and further details are given below in 2.9.

Physcomitrella patens ssp patens ecotype Gransden 2004 plants were grown on BCD medium with or without di-ammonium tartrate (Ashton et al., 1979).

BCD Medium	
Distilled Water	1 litre
$MgSO_4 \cdot 7H_20$ (or anhydrous $MgSO_4$)	250 mg (120 mg)
$KH2PO_4$	250 mg
KNO_3	1.01 g
$FeSO_4 \cdot 7H_20$	12.5 mg
Trace Element Solution	1 ml
4 M KOH to pH=6.5	
Agar	8 g
$CaCl_2$	1 mM

Marchantia polymorpha was grown on 1% (w/v) agar plates with 1/2 Gamborg B5 Basal salt, 0.5g/L MES (2-(*N*-morpholino)ethanesulfonic acid), pH 5.5 adjusted with KOH. Both bryophyte materials were grown under continuous illumination as used for *A. thaliana*.

2.8 Transformations

Competent *E. coli* cells were purchased from Bioline (α-select). The transformation method used for *E. coli* cells was based on the heat shock approach (Hanahan, 1983). 2.5 µl of plasmid or a recombination reaction was added into a 25 µl aliquot of competent cells thawed on ice removed from -80 °C freezer and incubated on ice for 30 minutes. The mixture was heat shocked at 42 °C for 45 seconds then put back on ice for another 2 minutes. 1 ml of LB medium was added and the culture was incubated at 37 °C for 1 hour shaking at 200 rpm. Afterwards, the culture was centrifuged for 30 seconds at 5,000 g and 800 µl supernatant discarded. Then the 200 µl of cell suspension was plated on LB agar with selection antibiotics. The plate was then incubated overnight at 37 °C.

Competent *A. tumefaciens* cells were prepared by ice-cold $CaCl_2$ solution treatment. For transformation, 4 µl of plasmid DNA was added into the frozen 25 µl aliquot and incubated at 37 °C for 5 minutes for the heat shock. Then 1 ml of LB was added and the culture was incubated at 28 °C for 2 - 4 hours shaking at 200 rpm. The rest of the process is the same as the transformation of *E. coli* cells, except the plate was incubated at 28 °C for 2 days.

The *A. thaliana* plants were transformed using the floral dip method with a few changes (Clough and Bent, 1998). *A. tumefaciens* cells were prepared in a 1 litre conical flask containing 400 ml of fresh LB medium with antibiotic selection until cell density reached saturation. Cells were then centrifuged at 1,000 g for 20 minutes and resuspended in 1 litre of infiltration medium (2.17 g/l half strength MS salts, 3.16 g/l full strength Gamborg B5 vitamins, 0.5 g/l MES, 50 g/l sucrose, 10 µg/l 6-benzylaminopurine). Just before dipping, 400 µl/l of Silwet L-77 was added. The above ground part of the plants was dipped in the *A. tumefaciens* solution for about 45 seconds with gentle agitation. The plants were then kept in normal conditions.

T1 plants were then selected on either soil subirrigated with 30 µg/mL BASTA (glufosinate ammonium; DHAI PROCIDA), or on MSO (Murashige and Skoog) medium with 0.8 % (w/v) phyto-agar petri dishes containing relevant antibiotics.

Polyethylene glycol-mediated transformation was used for *P. patens* (Schaefer et al., 1991, Cove, 2005). Protoplasts were acquired by using 1 % (w/v) Driselase in 0.5 M D-Mannitol solution to digest freshly grown around 7 days old protonemata and then pouring through filter papers and funnels. Calculated after using a hematocytometer, 300 µl of 1.2 x 10^6 ml^{-1} resuspended protoplasts were mixed with 15 µg of linearized plasmid DNA in 300 µl of PEG solution. Heat shock 5 min at 45 °C then leave at room temperature for 10 min. Plate on agar medium plates with sterile cellophane (Sigma-Aldrich) without any antibiotics. Then G418 and Hygromycin B (50 µg/ml) were used for selecting the transformants after no more than 7 days. The positive transformants were then confirmed by PCR. The transformation and following selection were done in collaboration with Dr Jörg Becker's group from the Instituto Gulbenkian de Ciência (IGC) in Portugal, mainly with the help of Dr Marcela H. Coronado and Dr Ann-Cathrin Lindner.

2.9 Microscopy

The pollen was stained with 4', 6-diamidino-2 phenylindole dihydrochloride (DAPI) for DNA visualizing (Park et al., 1998). All images were captured using the Nikon ECLIPSE 80i (Nikon, Japan) with an LED-based excitation source (CoolLED, presicExcite) and a

Plan Fluor 40x/1.3 NA oil immersion objective. Images were previewed, captured, and saved) in JPEG format using NIS-Elements Basic Research v3.0 software (Nikon, Japan).

For functional complementation analysis, an *A. thaliana* line with a GFP marker as in Borg *et al*., 2011 was used. The *duo1-1*$^{+/-}$ *pK7m34GW_promHTR10:H2B-GFP*$^{+/+}$ single locus marker line (i.e., having ~50 % GFP-positive wild type pollen grains) with a clear GFP signal in each pollen grain was generated from the *promHTR10:H2B-GFP*$^{+/-}$ lines. The progeny of this T1 single locus line with the highest apparent GFP signal was screened for heterozygous *duo1-1* and homozygous for the insertion marker. This homozygous marker line was maintained through screening previous generations and checking for homogenous GFP signal, as a means to prevent the silencing effect for the T-DNA insertion (Daxinger et al., 2008).

For analysis of complementation lines, mature pollen from T1 lines was checked by fluorescence microscopy. The first screen of the T1 lines was to identify the *duo1*$^{+/-}$ lines with a single insertion for the transgene (i.e., having ~50 % mCherry-positive pollen grains) and a reasonable mCherry signal. The frequency of tricellular (TCP) and bicellular (BCP) pollen grains was then scored using DAPI staining. The rescue efficiency of the cell cycle defect was calculated by the TCP frequency as a percentage in the population. The transactivation ability of the constructs was assessed by the frequency of the GFP-positive (GFP+) pollen grains. Similarly, the transactivation efficiency to the *HTR10* marker line was calculated by the GFP+ frequency as a percentage in the population.

For quantification of fluorescence, a pooled pollen sample from representative lines was analysed. The fluorescence of sperm cell nuclei was quantified in randomly selected pollen grains by image capture under standardized conditions. The exposure time was not pre-determined to avoid saturation. The images were then analysed using NIS-Elements. The total pixel intensity (TPI) of manually defined regions of interest encompassing sperm cell nuclei, with the cytoplasmic background subtracted, determined the true fluorescence of sperm cell nuclei (Borg et al., 2011). The average TPI of no less than 50 nuclei for each line were then calculated for statistically analysis.

2.10 Dual-luciferase reporter assays

Standard transient transformation of tobacco leaves was performed with modifications (Sparkes et al., 2006, Borg et al., 2011). Infiltrated leaves were excised with a 9 mm cork borer and ground in 500 mL of 1x Passive Lysis Buffer (Promega). Leaf extracts were centrifuged at 4°C at 15,100 g to pellet debris. The firefly luciferase assay buffer (25 mM glycylglycine, 15 mM KPO_4, pH 8.0, 4 mM EGTA, 2 mM ATP, 1 mM DTT, 15 mM $MgSO_4$, 0.1 mM CoA, and 75 mM luciferin with final pH adjusted to 8.0) and *Renilla* luciferase assay buffer (1.1 M NaCl, 2.2 mM Na_2EDTA, 0.22 M KPO4, pH 5.1, 0.44 mg/mL BSA, and 1.43 mM coelenterazine with final pH adjusted to 5.0) (Dyer et al., 2000) were prepared before the reading. Two 25 mL aliquots were separately assayed with 200 mL of each assay buffer. Relative luciferase activity (FLuc/RLuc) was calculated for each infiltration.

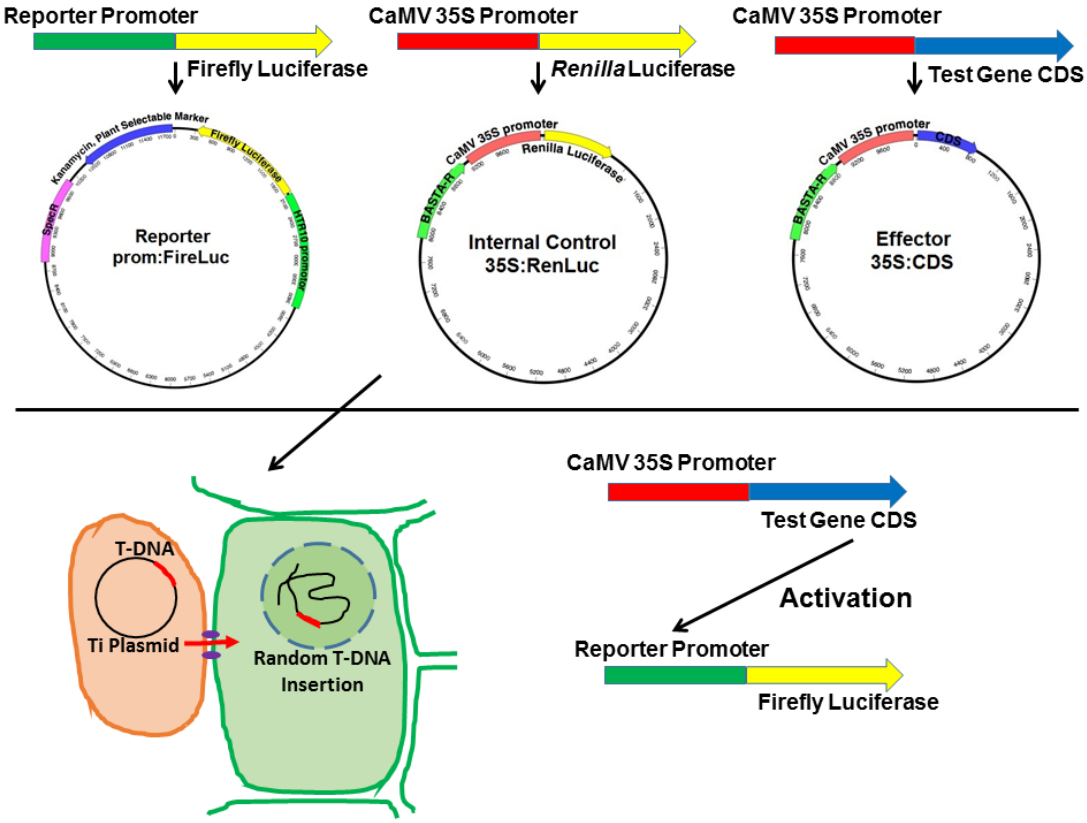


Figure 2. 1 Diagram of the dual-luciferase transient assay.
Three types of constructs are used: reporters, effectors, and an internal control. The gene of interest (blue) is driven by the CaMV 35S promoter (red), which will be expressed in the tobacco leaf cells. The interaction between the effector and the potential target promoter (green) is correlated with the measured level of the firefly luciferase. The internal control is the *Renilla* luciferase driven by the 35S promoter that acts as a benchmark for the effector expression level. They are then delivered through Agrobacteria. Diagram adapted from Păcurar et al. (2011).

2.11 Mammalian two-hybrid system

Some modifications were made on the polyethylenimine (PEI) (Sigma)-mediated transient transfections of the HEK 293T cells (Thermo Fisher Scientific) (Itoh et al., 2015). A total of 0.5 µg DNA (0.18 µg β-gal plasmid, 0.23 µg firefly luciferase plasmid, 0.1 µg effector plasmid) was mixed with 25 µl PBS (Sigma), making up to 200 µl with 25 µl of 0.5 mM PEI solution in PBS, 150 µl of DMEM and 10 % FBS of the transfection mixture. After incubating for 36 hours, the cells were lysed and measured for firefly luminescence as in the section 2.10. The β-gal levels were measured using 100 µl of substrate (10 ml contains 6 ml 0.1 M Na_2HPO_4, 0.1 ml 1 M KCl , 0.1 ml 0.1M $MgCl_2$, 4 ml 0.1 M NaH_2PO_4, 20 mg ONPG, 35 µl Mercaptoethanol) for normalisation (FLuc/β-gal).

2.12 Statistics analysis

All statistic tests were performed by using Microsoft Excel or IBM SPSS Statistics. Statistical assessment of scoring data was performed with a χ^2 test. Comparison of the means between groups of data was done with t-test after checking for Student's t-distribution. All tests were two-sided with statistically significant outcomes determined using a level of 0.05. Unless specified, all error bars in the charts represent the standard error (SE).

2.13 Phylogenetic analysis

Sequence alignments and phylogenetic trees were constructed using MEGA 7.0 (Kumar et al., 2016). Unless specified, the default alignment was based on the protein sequences using ClustalW (Higgins and Sharp, 1988). The default phylogenetic tree was based on the protein coding nucleotide sequences, using the maximum likelihood method and the Tamura-Nei model (Tamura and Nei, 1993). Distance-based methods such as neighbor-joining and maximum parsimony were also frequently used to quickly generate an initial tree for parameter testing or further heuristic tree-searching. The default bootstrap number was 1000. Positive selection tests were performed using PAML 4.9 (Yang, 2007), with either Branch-Site Model or Clade Model.

Chapter 3 Analysis of DUO1 Phylogeny and Sequence Conservation

Abstract

DUO1 is a key regulator of the male germline development in plants and controls the cell division and cell differentiation. The closest relatives of DUO1 are members of the GAMYB-like protein family. The similarities and differences between the two clades are important for the better understanding of the origin of DUO1. Evidence presented in this chapter revealed that DUO1 first evolved in the green algae group Charophyta, and the time of its emergence coincided with the beginning of sexual reproduction in the Streptophyta clade. Although multiple whole genome duplication events have happened in a range of land plant lineages like Brassicaceae, DUO1 has been diploidised in most species, which means there is only one copy in a species. Here a loss of selective pressure and the subsequent loss-of-function mutation were demonstrated in some species where multiple copies of DUO1 exist.

3.1 Introduction

As discussed in Chapter 1, DUO1 exists in all land plants. It is a R2R3 MYB protein and has been confirmed to be responsible for the activation of many genes with a MYB binding site in their promoter regions (Rotman et al., 2005, Brownfield et al., 2009a, Borg et al., 2011). Compared to the other R2R3 MYB proteins, DUO1 has a supernumerary lysine (K66) within the R3 repeat. DUO1 homologs in all angiosperm species examined in these studies were found to possess this lysine, but no other R2R3 MYB proteins were found to have this feature (Rotman et al., 2005, Brownfield et al., 2009a). The closest sister clade to DUO1 is the GAMYB clade. Both clades are related to male reproduction.

MYB domains are known to bind to double-stranded DNA (Solano et al., 1995, Solano et al., 1997, Gubler et al., 1999). The Gibberellin Response Element (GARE) sequence is TAACAAA, which was first discovered in barley (Skriver et al., 1991, Gubler et al., 1995, Gubler et al., 1999). In *A. thaliana*, this sequence has been found in GAMYB target gene promoters and MYB97, MYB101, and MYB120 have been reported to bind the same sequence in the electrophoretic mobility shift assays (EMSAs) (Liang et al., 2013). On the other hand, DUO1 has a distinct binding sequence of A/TAACCGT/C and disruption of two copies of this motif in the *HTR10* (histone H3.10) promoter almost eliminated the male germline-specific transcription of this promoter *in planta* (Borg et al., 2011).

However, previous studies had not looked into the DUO1 homologs extensively across the whole plant kingdom. There was little knowledge regarding the origin of DUO1. The relation, similarities, and differences between the DUO1 and GAMYB proteins were not described in details, either. This chapter will go into the details of sequence variation among DUO1 homologs, and their complicated connections with the GAMYB clade.

3.2 Sequence analysis of land plants DUO1 homologs

A total of 85 nucleotide sequences of *DUO1* homologs were compiled from 52 land plant species from different groups. The species and the copy numbers in each species are presented in a cladogram and discussed later (see Section 3.5). The data were collected from Phytozome (phytozome.jgi.doe.gov), Ensembl Plants (plants.ensembl.org), NCBI

(www.ncbi.nlm.nih.gov), and various sequencing projects (Albert et al., 2013, Nystedt et al., 2013).

A typical *DUO1* locus contains three exons and two introns (Figure 3. 1). The first intron is located within the R2R3 MYB domain. The second intron is usually in a position with highly variable sequences between species. However, the intron length can vary widely from less than 100 bp in *Arabidopsis thaliana* to more than 1 kb in *Sphagum fallax*.

Some sequences were found to be incorrectly annotated (e.g. *Selaginella moellendorffii*) and improved gene-model predictions were made based on the GT-AG splicing site (Sharp and Burge, 1997). These predictions mainly required determining three critical parts of the sequence: the start codon ATG, the first intron, and the second intron.

The start codon was identified as the closest upstream ATG that produced the correct open reading frame (ORF) for the R2 repeat. The first intron was identified by the GT-AG splicing site and confirmed by the integrity of the R2R3 MYB domain. The most challenging prediction was for the position of the second intron. The first step was to look for the donor site (GT) near the first stop codon that breaks the ORF in the genomic sequence; the second step was to get all the downstream accepter site (AG) that produce a correct ORF, usually indicated by the C-terminal homologous protein sequence in DUO1 homologs; the final step was to exam all possible accepter site by checking for the polypyrimidine tract and the branch site (Black, 2003).

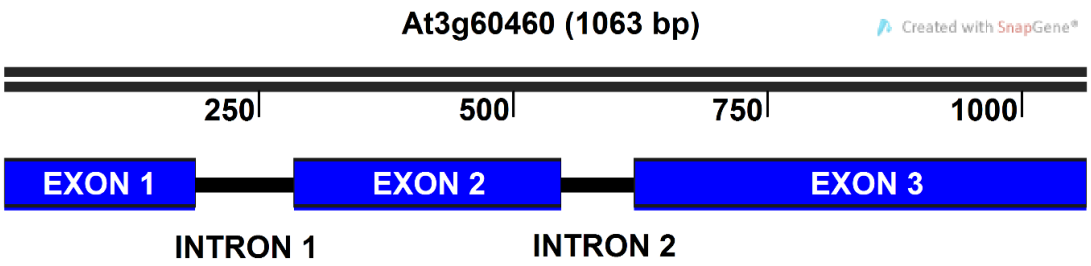


Figure 3. 1 Structure of DUO1 locus in *A. thaliana* (At3g60460).
At3g60460 represents a typical DUO1 locus in all land plants, which contains three exons (894 bp in total) and two introns (169 bp in total). The three exons are 188 bp, 262 bp, and 444 bp (excluding stop codon), respectively. The two introns are 97 bp and 72 bp, respectively. The diagram was created using SnapGene®.

Protein Lengths of DUO1 Homologs

Angiosperms					
AthDUO1	SlyDUO1A	SlyDUO1B	OsaDUO1	MacDUO1	AtrDUO1
297	319	304	343	322	288
Pre-angiosperms					
PabDUO1	SmoDUO1A	SmoDUO1B	PpaDUO1A	PpaDUO1B	MpoDUO1
437	373	456	493	489	402

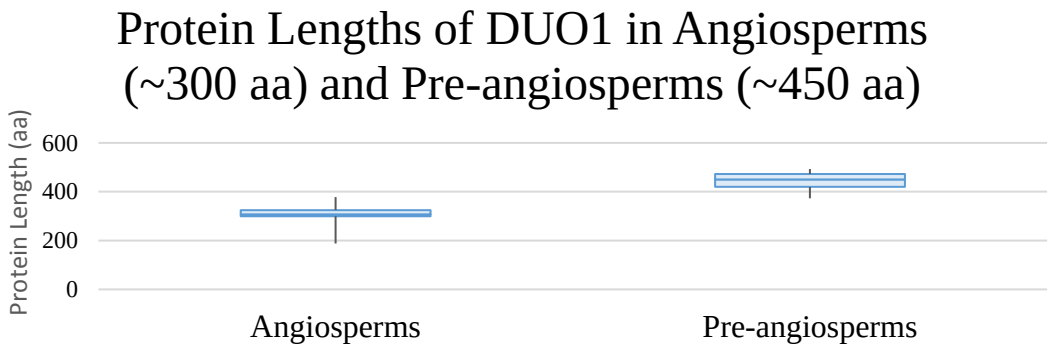


Figure 3. 2 DUO1 protein length variation in different land plants.
Some of the sequences with typical protein lengths from both angiosperms and pre-angiosperms were displayed here. The box and whisker plot included 88 sequences collected in total. Ath=*Arabidopsis thaliana*, Osa=*Oryza sativa*, Sly=*Solanum lycopersicum*, Atr=*Amborella trichopoda*, Pab=*Picea abies*, Smo=*Selaginella moellendorffii*, Ppa=*Physcomitrella patens*, Mpo=*Marchantia polymorpha*.

Once the CDS of the *DUO1* homologs were predicted, the amino acid sequences could be analysed. The lengths of the complete DUO1 proteins showed variation between different species and clades (Figure 3. 2). Pre-angiosperm DUO1 proteins appear to be distinctively longer, ranging from 373 to 493 amino acids long, in comparison to angiosperm DUO1 homologs, which are around 300 amino acids and have a narrower range (288-343 aa).

After all DUO1 sequences were aligned, it became clear that there are two conserved regions present within land plants (Figure 3. 3). The highly conserved MYB DNA-binding domain and a region of lower conservation at or near the C-terminus, characterized by the conserved DxFD motif.

As discussed in Chapter 1, MYB domains are well studied and known for their astonishing conservation. The DUO1 MYB domain is no different. Not only does it have high similarities between species, it also resembles other R2R3 MYB domains, which is discussed later in this Chapter. The proportion of acidic residues at the C-terminal end increased from only 16 % in bryophytes to 21-32 % in eudicots. Detailed sequence and functional analysis of the C-terminal region is presented in Chapter 5.

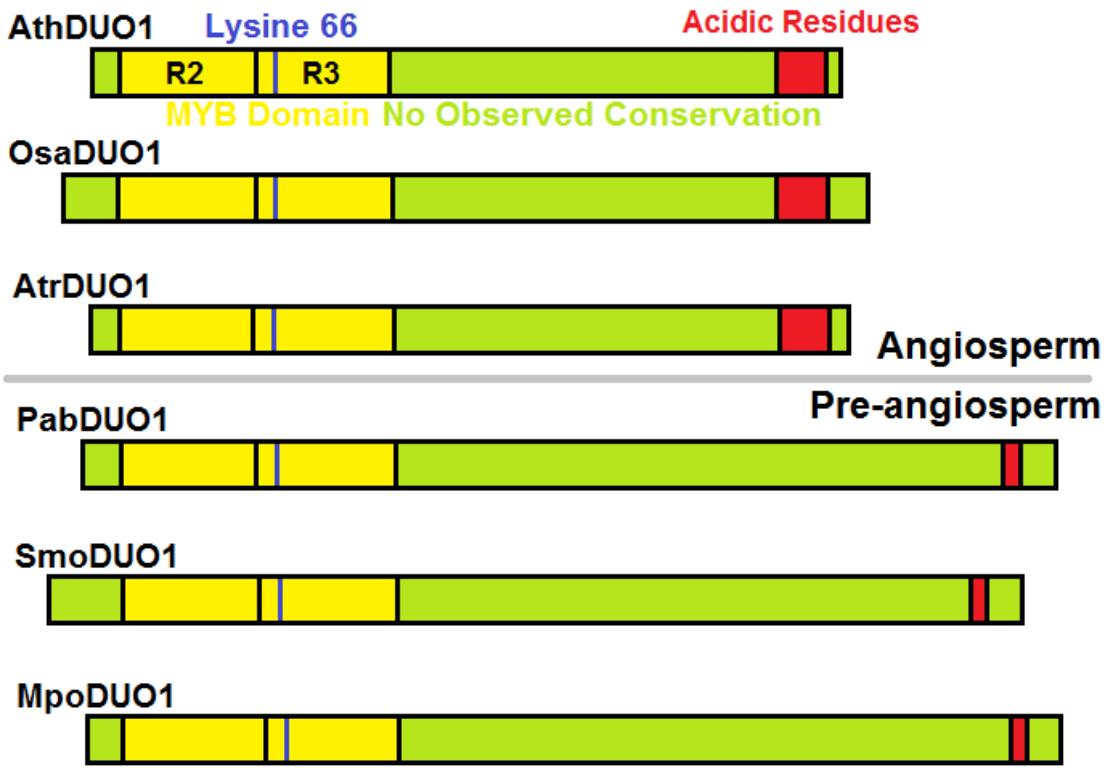


Figure 3. 3 Domain structure of DUO1 homologs from different land plant species. DUO1 proteins are shown for six representative species of major land plant clades, including eudicot (Ath=*Arabidopsis thaliana*), monocot (Osa=*Oryza sativa*), basal angiosperm (Atr=*Amborella trichopoda*), gymnosperm (Pab=*Picea abies*), lycophyte (Smo=*Selaginella moellendorffii*), and bryophyte clades (Mpo=*Marchantia polymorpha*).

The initial observation suggested that both the MYB domain and C-terminus motif are in relatively fixed positions at either end. Further calculations demonstrate that the length of the region linking the MYB and the C-terminus is strongly correlated with the total length of the protein (Figure 3. 4). Thus it is reasonable to say that the length variations among DUO1 proteins are the results of the linking region length variations. This implies a possible detrimental effect of changes at either end. However, the fusion with different florescent proteins at both ends have been frequently used for the studies of DUO1. It is unclear whether this has caused the failure of expression in certain florescent protein fused constructs (mentioned in Section 1.2). Most angiosperm DUO1 homologs tend to have a total protein length around 280-350 amino acids long, and the linking region is around 160-200. The limited range of DUO1 protein length in angiosperms suggests that there may be greater functional constraints on protein length in angiosperms compared with pre-angiosperms.

Species	Gene name	Protein Length	Linking Region Length
Picea abies	PabDUO1	437	274
Marchantia polymorpha	MpoDUO1	402	266
Sorghum bicolor	SbiDUO1A	378	219
Ananas comosus	AcoDUO1	349	184
Oryza sativa	OsaDUO1	343	180
Zea mays	ZmaDUO1A	339	191
Lilium longiflorum	LloDUO1	325	191
Solanum lycopersicum	SlyDUO1A	319	198
Solanum lycopersicum	SlyDUO1B	304	165
Mimulus guttatus	MguDUO1B	298	167
Arabidopsis thaliana	AthDUO1	297	175
Amborella trichopoda	AtrDUO1	288	162
Carica papaya	CpaDUO1	265	139
Spirodela polyrhiza	SpoDUO1	242	111
Mimulus guttatus	MguDUO1A	236	116

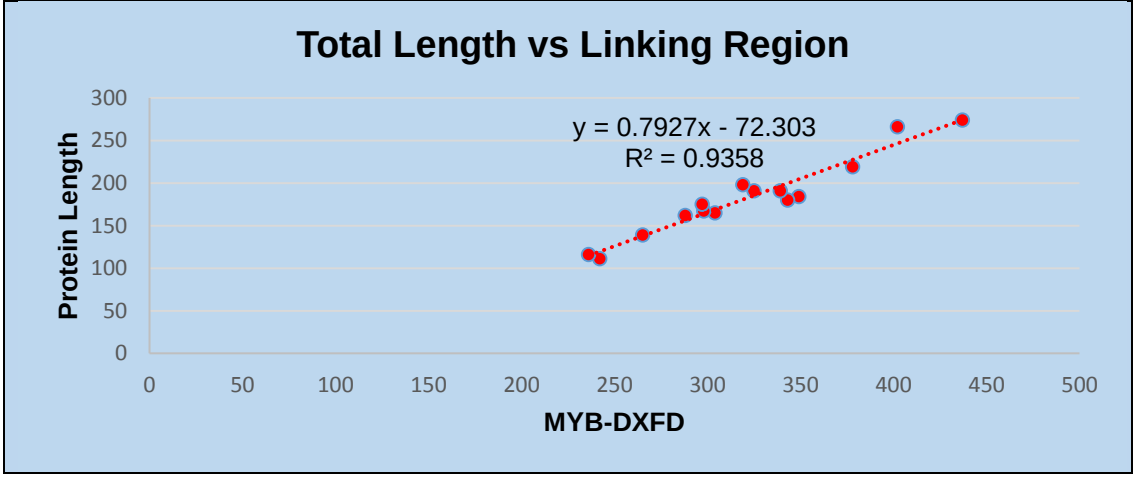


Figure 3. 4 The MYB domain and C-terminus motif are located in conserved positions, the length of the linking region in-between determines the length of a DUO1 homolog. A total of 15 sequences from 13 Species was listed ranging from just 236 amino acids long (MguDUO1A) to 378 amino acids (SbiDUO1A, anigosperm), and 437 amino acids (PabDUO1, pre-angiosperm). The chart shows that the correlation between the protein total lengths and the lengths of the linking region (calculated from the end of the MYB domain to the DXFD motif) are strongly correlated ($R^2 = 0.9358$). According to the Gaussian correlation inequality (GCI) Theorem, we can expect most angiosperms have a DUO1 homolog 280-350 amino acids long with the linking region around 160-200 amino acids. The ones that fall outside of these ranges may have risen as a result of evolution towards a special environment, or degeneration due to relaxed selection pressure (e.g. post duplication).

There is no obvious conservation in sequence in the linking region. That is not to say this part of the protein is not important or can be replaced with any sequences. Some secondary structure predictions were made to search for any significant details.

First the hydrophobicity of different homologs was studied (Figure 3. 5). The hydropathicity (relative hydrophobicity) profiles were made using ExPASy (Kyte and Doolittle, 1982, Wilkins et al., 1999). These homologs, from *A. thaliana*, *O. sativa*, *L. longiflorum*, *M. polymorpha*, and *S. lycopersicum* were functionally tested (see Chapter 4). There was no clear patterns in the diagrams that were shared by these proteins.

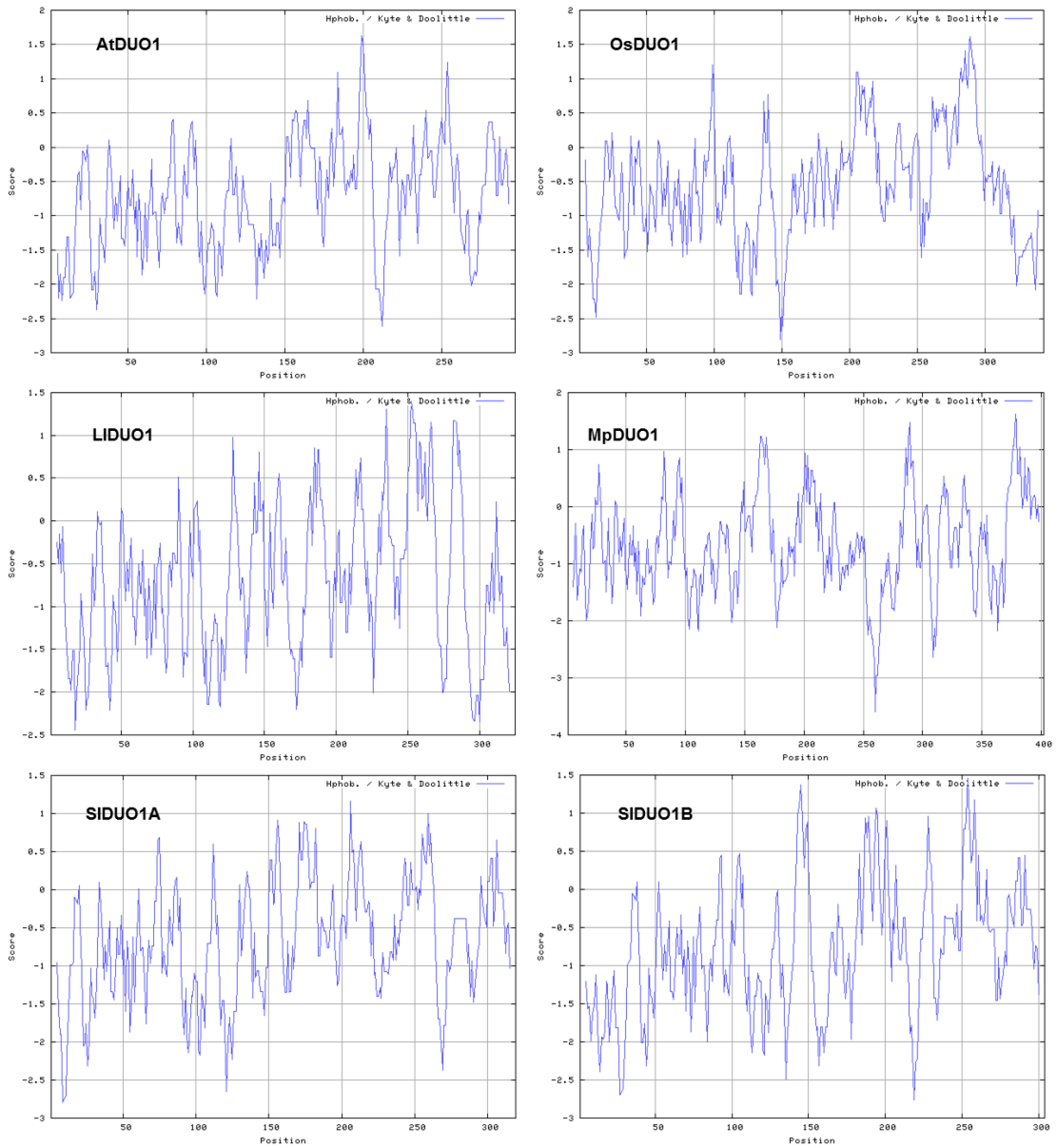


Figure 3. 5 Hydropathicity profiles of various DUO1 homologs.
There is no observable patterns of the DUO1 homologs. The peaks and valleys seem to be a bit tighter around the MYB region, and apart from MpDUO1, the hydropathicity of the other homologs before the C-terminal end reaches a low level. However, the significance of this feature is unclear.

Another feature that was looked into is shown in Figure 3. 6. The secondary structures of AtDUO1 were predicted using SPIDER2 (Heffernan et al., 2016). For all the predictions that were done by SPIDER2 in this work, the letters used are all in the Dictionary of Protein Secondary Structure (DSSP) classification:

- G = 3-turn helix (3_{10} helix). Min length 3 residues.
- H = 4-turn helix (α helix). Min length 4 residues.
- I = 5-turn helix (π helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn).
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation).
- S = bend (the only non-hydrogen-bond based assignment).
- C = coil (residues which are not in any of the above conformations).

The "H" indicates high probability of residues that might be in a helix. However, it requires a minimum of 3 residues to form a 3_{10} helix, and 4 for α-helix, thus we do not need to consider the residues to be in a helical structure if the number of such connecting residues are only two or less, without any potentially connecting residues in the vicinity. On the other hand, the closely interspersed residues with high probabilities of being in a helix are likely forming one, such is the case of the third α-helix of the R2 repeat.

Therefore, apart from the six α-helices in the MYB region, only one helix was predicted for the DUO1 secondary structure. This one helix inside the linking region was not present in OsDUO1 or SlDUO1A, thus unlikely being important in terms of evolution.

The blue numbers indicate residues that have an extremely low relative accessible surface area (rASA). These values give information about the protein secondary and tertiary structure (Momen-Roknabadi et al., 2008). It is very intriguing that in the conserved C-terminal region, the F279, L280, F283, and F288 all have a low rASA score, which could have the potential to form a helical structure. As mentioned earlier, the residues are relatively conserved among angiosperms, which leads to the similarities of the rASA values in the C-terminus. However, the secondary structure predictions for some DUO1 proteins contain a helix structure in this region (see Chapter 4).

The linking region in this regard has not shown any recognisable patterns for the rASA. The sequence analysis to predict any functional conservation for this diverse region so far has not yet yielded any conclusive results. It is hard to imagine a conserved tertiary structure here given the diversity of the sequences and secondary structures among different species. The significance of the linking region therefore can only be determined by a more direct method (See Chapter 5).

SPIDER2 result for AtDUO1

The Predicted Secondary Structures for AtDUO1:

```
SEQ : 1   MEAKKEEIKKGPWKAEEDEVLINHVKRYGPRDWSSIRSKGLLQRTGKSCR 50
SS  : 1   ----H---------HHHHHHHHHHHHHH----HHHHHH------H-H-H-  50
rASA: 1   78466653563415461053015106631462142037523253125313 50

SEQ : 51  LRWVNKLRPNLKNGCKFSADEERTVIELQSEFGNKWARIATYLPGRTDND 100
SS  : 51  --HH-H------------HHHHHHHHHHHHHH--HHHHHHHH------HH 100
rASA: 51  23122313442655360445115101401552534115005415633443 100

SEQ : 101 VKNFWSSRQKRLARILHNSSDASSSSFNPKSSSSHRLKGKNVKPIRQSSQ 150
SS  : 101 HHHHHHHHHHHHHHHH--------------------------------- 150
rASA: 101 14421343355336456878767677656766665647777565476666 150

SEQ : 151 GFGLVEEEVTVSSSCSQMVPYSSDQVGDEVLRLPDLGVKLEHQPFAFGTD 200
SS  : 151 -------------------------------------------------- 200
rASA: 151 64554666566676577756466765565436356354645445443655 200

SEQ : 201 LVLAEYSDSQNDANQQAISPFSPESRELLARLDDPFYYDILGPADSSEPL 250
SS  : 201 ------------------------HHHHH-------HH----------- 250
rASA: 201 34456356656555465444234625511551733212321256646544 250

SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLDDFPADMFDQVDPIPSP 297
SS  : 251 ----------------------------H----------------- 297
rASA: 251 45354432544634563663566653421133134521543564675 297
```

Figure 3. 6 Secondary structure prediction of AtDUO1 using SPIDER2.
There is likely a helix in the MYB domain where the "H"s were broken up, making it six α-helices in the whole MYB region. This agrees with the literature studies on MYB proteins. There is one helix in the linking region, and no predicted structures in the C-terminal end. rASA, the relative ASA [0,9]; Buried residues with rASA <20% are labelled blue; in the Dictionary of Protein Secondary Structure (DSSP) classification, H = 4-turn helix (α helix). Min length 4 residues.

3.3 Protein model of the DUO1 MYB domain

The conservation of the MYB domains can be seen in their identity and similarity scores (Table 3. 1). However, the identity scores drop quickly with rest of the sequence, from ~22 % between AthDUO1 and OsaDUO1 to only ~9 % between AthDUO1 and MpoDUO1.

Table 3. 1 Similarity (Identity) scores of the DUO1 MYB domains in land plants.

Sim/Ide (%)	AthDUO1	OsaDUO1	AtrDUO1	PabDUO1	SmoDUO1A	MpoDUO1A
AthDUO1	**100**					
OsaDUO1	**89.5 (83.7)**	**100**				
AtrDUO1	**87.6 (79.8)**	**91.4 (86.5)**	**100**			
PabDUO1	**86.7 (75.0)**	**89.5 (79.8)**	**95.2 (87.5)**	**100**		
SmoDUO1A	**86.7 (76.0)**	**87.6 (78.8)**	**93.3 (86.5)**	**98.1 (97.1)**	**100**	
MpoDUO1A	**83.8 (71.1)**	**86.7 (76.9)**	**92.4 (84.6)**	**97.1 (96.2)**	**97.1 (95.2)**	**100**

Ath=*Arabidopsis thaliana*, Osa=*Oryza sativa*, Atr=*Amborella trichopoda*, Pab=*Picea abies*, Smo=*Selaginella moellendorffii*, Mpo=*Marchantia polymorpha*.

Although mutation experiments have been done for the AtDUO1 at the amino acid level, yet little is known about its protein structure. In order to have a better understanding about how DUO1 is binding to DNA and where the Lysine 66 is, it is important to be able to visualise this protein.

MYB protein family is well studied. Even though there is no close plant MYB protein structure that has been built through X-ray or NMR, other MYBs from human or mouse are already documented in the Protein Data Bank (Berman et al., 2000). A protein structure prediction based on an existing template from the PDB can be conducted using the website SWISS-Model (*swissmodel.expasy.org*) (Schwede et al., 2003, Arnold et al., 2006, Berman, 2008, Berman et al., 2014).

Since the MYB domain of DUO1 is much more conserved than the rest of the sequence, and it was clear that the MYB domain was the DNA binding region. So to use only the MYB domain was more likely to result in a reliable model that contains the DNA binding information.

The protein sequence used was the *Arabidopsis thaliana* DUO1 R2R3 MYB domain, which starts from the 8th to the 112th amino acid (Rotman et al., 2005). The sequence was then uploaded to the SWISS-Model website and the search for the templates was performed automatically using BLAST under the framework based on PERL (Berman et al., 2000, Kiefer et al., 2009).

Two templates stood out from dozens of potential templates with a relatively higher GQEM score. Their SMTL ids are 1h8a.1.C (identity 48.51%, similarity 58.65%) and 1gv2.1.A. The former one was from human and the latter one from mouse. However, only the 1h8a.1.C was in a DNA binding status (Tahirov et al., 2002), which was also expected for DUO1, so the model based on the first template was the preferable one.

Two different models with the same template were built using two different alignments (Figure 3. 7). The PDB format files were then downloaded and imaged in the software PyMOL (DeLano, 2004, Labby, 2013). Using PyMOL aligning Model 1 to the template, 102 residues were aligned, and the RMS (Root-mean-square) was 0.051 (90 to 90 atoms). For Model 2 aligning to the template, 102 residues were aligned, the RMS was 0.054 (89 to 89 atoms). When the two models aligned to each other, for all 104 aligned residues, the RMS was 0.005 (81 to 81 atoms). Although the change was small, it concerned a major change in the secondary structure prediction. Thus it was worth determining which model was more probable for further functional studies. However, the true structure of the DUO1 MYB domain cannot be confirmed until a more direct approach is made, instead of predictions.

Various tools were used for scoring the two models. On SWISS-MODEL, Model 1 had GMQE of 0.80 and QMEAN4 of -2.06; Model 2 had GMQE of 0.78 and QMEAN4 of -2.49 (See Appendix). On ProSA (Wiederstein and Sippl, 2007), the Model 1 Z-Score was -5.87; and Model 2 was -5.98. On ERRAT (Colovos and Yeates, 1993), the Model 1 overall quality factor was 83.333; Model 2 was 72.917. It showed the 9 residues that were supposed to be the second helix in Model 2 had the error values between 95%-99%; in comparison to only 2 residues in Model 1 at this region.

These analyses showed that although the two models were very similar, Model 1 seems to perform slightly better in scores. However, the R2 repeat of this model only displayed two α-helixes, which would disrupt the hydrophobic core and the binding ability. This defies the well-established model of the MYB domain.

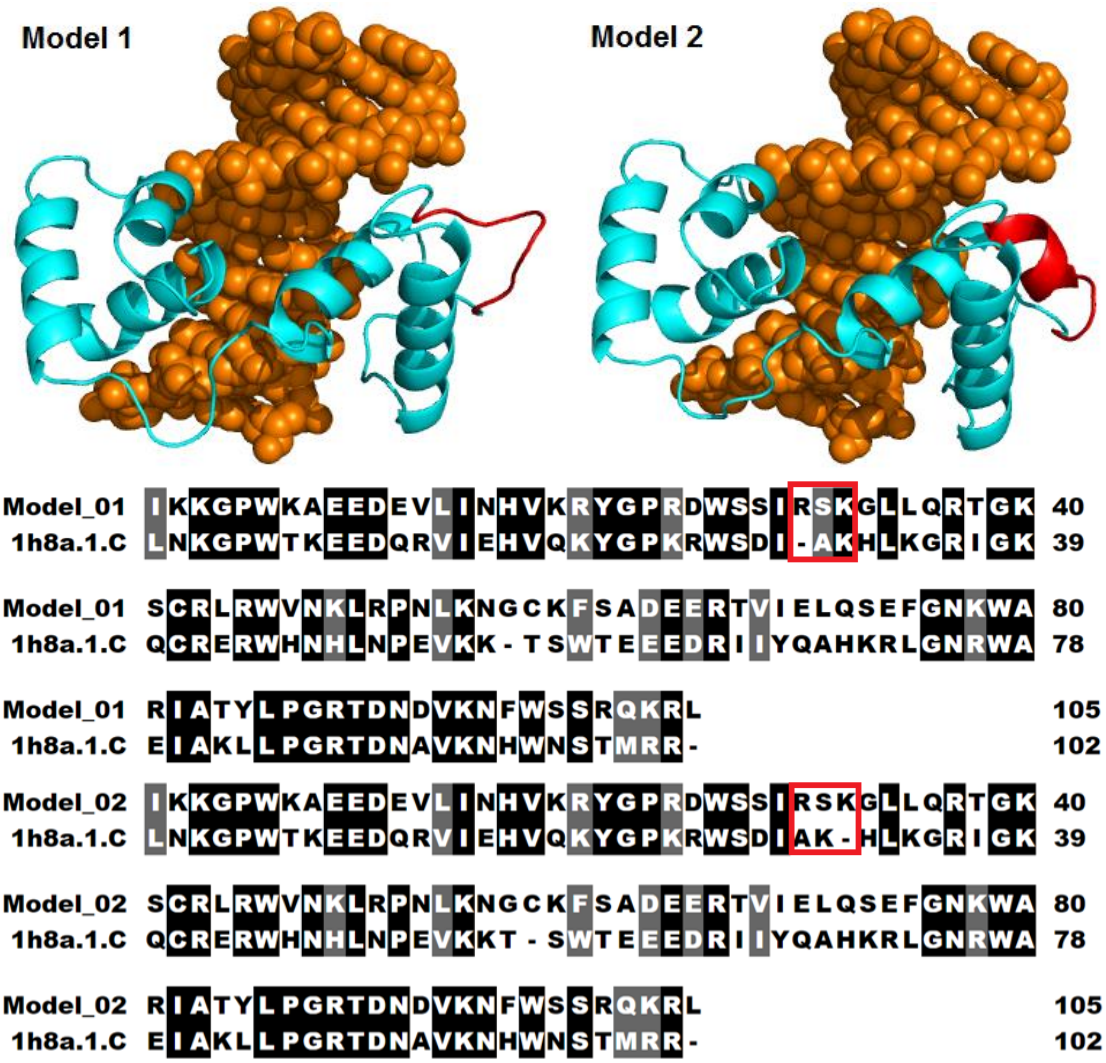


Figure 3. 7 The two protein models for the MYB domain of *A. thaliana* DUO1.
Both models are built based on the template 1h8a.1.C. For Model 1, there are three α-helices in the R2 repeat and three in the R3; for Model 2, both repeats have three α-helices. The structure marked red is where the two models differ from each other. This is caused by the alignment difference of the arginine residue in the red box. This insertion breaks the missing helix in the prediction. Since the MYB domain is well known for having three α-helices that forms a hydrophobic core, the second model will be adopted in this thesis. Assuming DUO1 MYB behaves similar to the template 1h8a.1.C, the last helix of each repeat interacts with the major groove of the DNA (Tahirov et al., 2002).

Getting a closer look at the differences in the two alignments with their models, it revealed that Model 1 had an insertion in the missing helix, which was the arginine aligned to the alanine highlighted by the red box in Figure 3. 7 (R → A). Since there was no good reason to "squeeze" this residue in such an unfavourable conformation, and there was also a secondary structure prediction of an α-helix in this position (Figure 3. 4), Model 2 and its alignment are more likely to be the correct prediction.

In this thesis, Model 2 was regarded as the accurate prediction. This means all the analysis that based on the MYB protein structure assumed that there were three α-helices in both the R2 and R3 repeat.

3.4 The origin and divergence of DUO1 and GAMYB family

The DUO1 clade is closely related to the GAMYB clade. The MYB domains of these proteins show a high level of similarity (Figure 3. 8). During the research, three regions in the MYB domains of the two clades displayed significant differences from each other. They were consequently labelled A, B, and C by Dr Tomokazu Kawashima. The first region (A) in DUO1 forms a miR159 binding site in DUO1. It is located in the second helix of the R2 repeat. Region B contains a signature supernumerary lysine (K66), which is only present in DUO1 homologs. It is a part of sequence that links the R2 and R3 repeat. Region C forms the third helix of the R3 repeat, which is responsible for the target DNA interaction.

To trace the origin of the *DUO1* gene, the point where the *DUO1* and *GAMYB* clades separate from each other is crucial. The sister clade to all land plants is the streptophyte, Charophyta. The homologs from this group have proven to be very informative. The alignment of these homologs with the land plant DUO1 and GAMYB proteins at the A, B, and C region was shown in Figure 3. 8. A *Spirogyra* DUO1 and a *Chara* DUO1 homolog have the signature lysine in region B, but the closest *Klebsormidum* sequence does not. This suggests that *DUO1* homologs first emerged in Charophyta after the separation of Zygnematophyceae, Charophyceae, and Klebsormidiophyceae. Interestingly, the *Spirogyra* sequence of DUO1 shows signs of degeneration, but the GAMYB proteins (which also related to male reproduction) of a *Spirogyra* species acquired from Blast4OneKP (db.cngb.org/blast4onekp) showed no sign of such degeneration in the MYB domain (alignment see Appendix).

As discussed in Section 1.8, traditionally it is considered that the land plants lineage started from a group of Charales, Coleochaetales, or some members of the Zygnematophyceae. Recent evidence all supports the claim that the Zygnematophyceae, or a clade that also including the Coleochaetales lineage, is the closest relative.

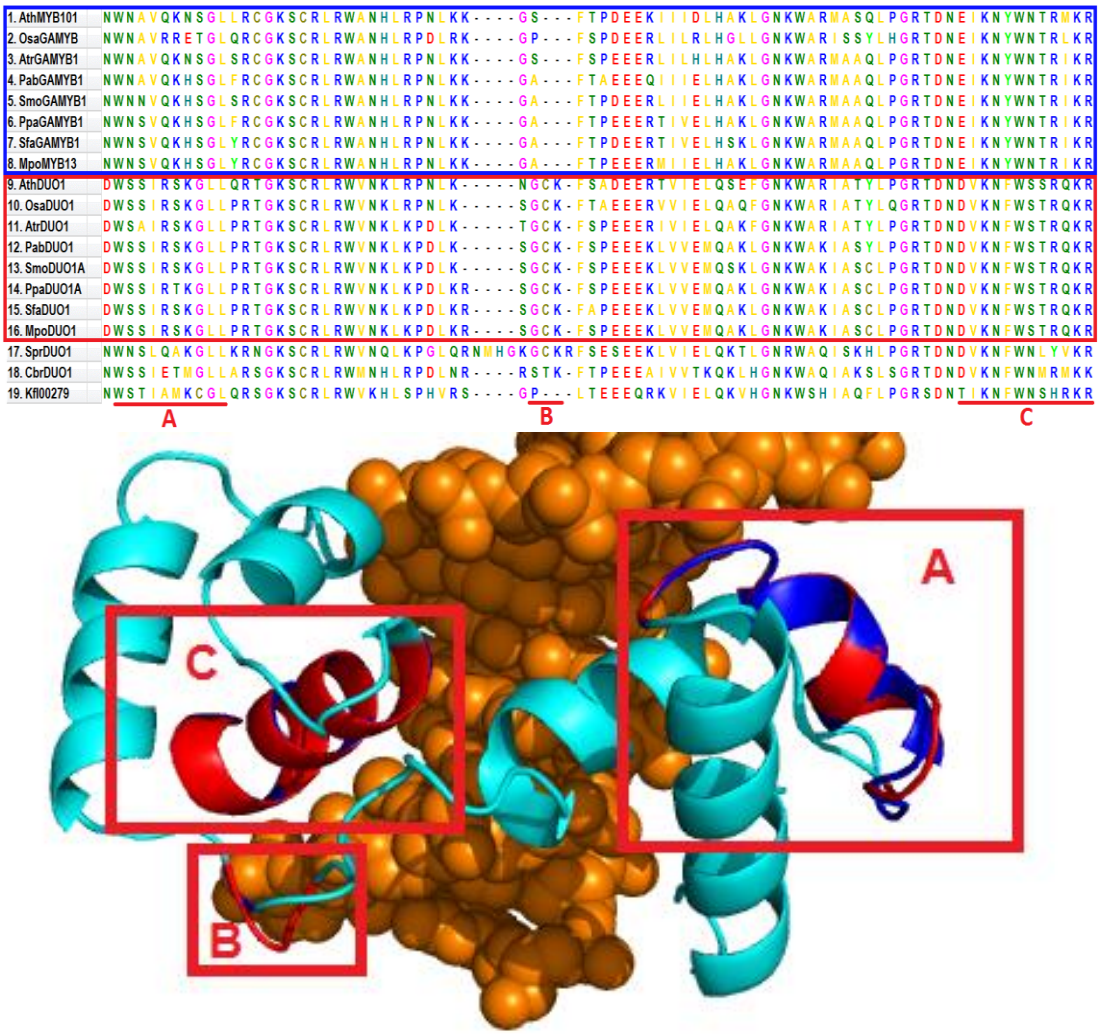


Figure 3. 8 Alignment of the DUO1 and GAMYB MYB domains.
The MYB binding domain of DUO1 is highly conserved and distinguish itself from its GAMYB relatives mainly in three regions (marked A, B, and C). The A region is changed in DUO1 to form a miR159 binding site. While the C region is regarded as DNA recognition site, the B region contains the DUO1 signature residue, Lysine 66 (K66). Homologs form *Spirogyra panensis* and *Chara braunii* bear this feature, despite some clear differences from the land plant homologs. However, the closest *Klebsormidum* homolog do not have K66, thus cannot be considered as DUO1 sequences. The protein model is an overlap of the MYB domains from the MpDUO1 and MpMYB13 (Mapoly1089s0002.1). The three regions are coloured and highlighted. MpDUO1 is shown in red and MpMYB13 is shown in blue. Ath=*Arabidopsis thaliana*, Osa=*Oryza sativa*, Atr=*Amborella trichopoda*, Pab=*Picea abies*, Smo=*Selaginella moellendorffii*, Ppa=*Physcomitrella patens*, Sfa=*Sphagum fallax*, Mpo=*Marchantia polymorpha*, Spr=*Spirogyra pranensis*, Cbr=*Chara braunii*, Kfl=*Klebsormidium flaccidum*.

Here we used the genomes of *Spirogyra pranensis* and *Closterium peracerosum* (Zygnematophyceae), *Chara braunii* (Charophyceae), and *Klebsormidium flaccidum* (Klebsormidiophyceae) to search for candidates for DUO1 homologs. Then DUO1 and GAMYB proteins from various land plant species were selected, including bryophyte, lycophyte, gymnosperm, basal angiosperm, monocot, and eudicot clades. The

phylogenetic tree was rooted using the DUO1 clade to search for early divergent homologous sequences from algal species (Figure 3. 9). The result confirmed the finding from the alignment in Figure 3. 8.

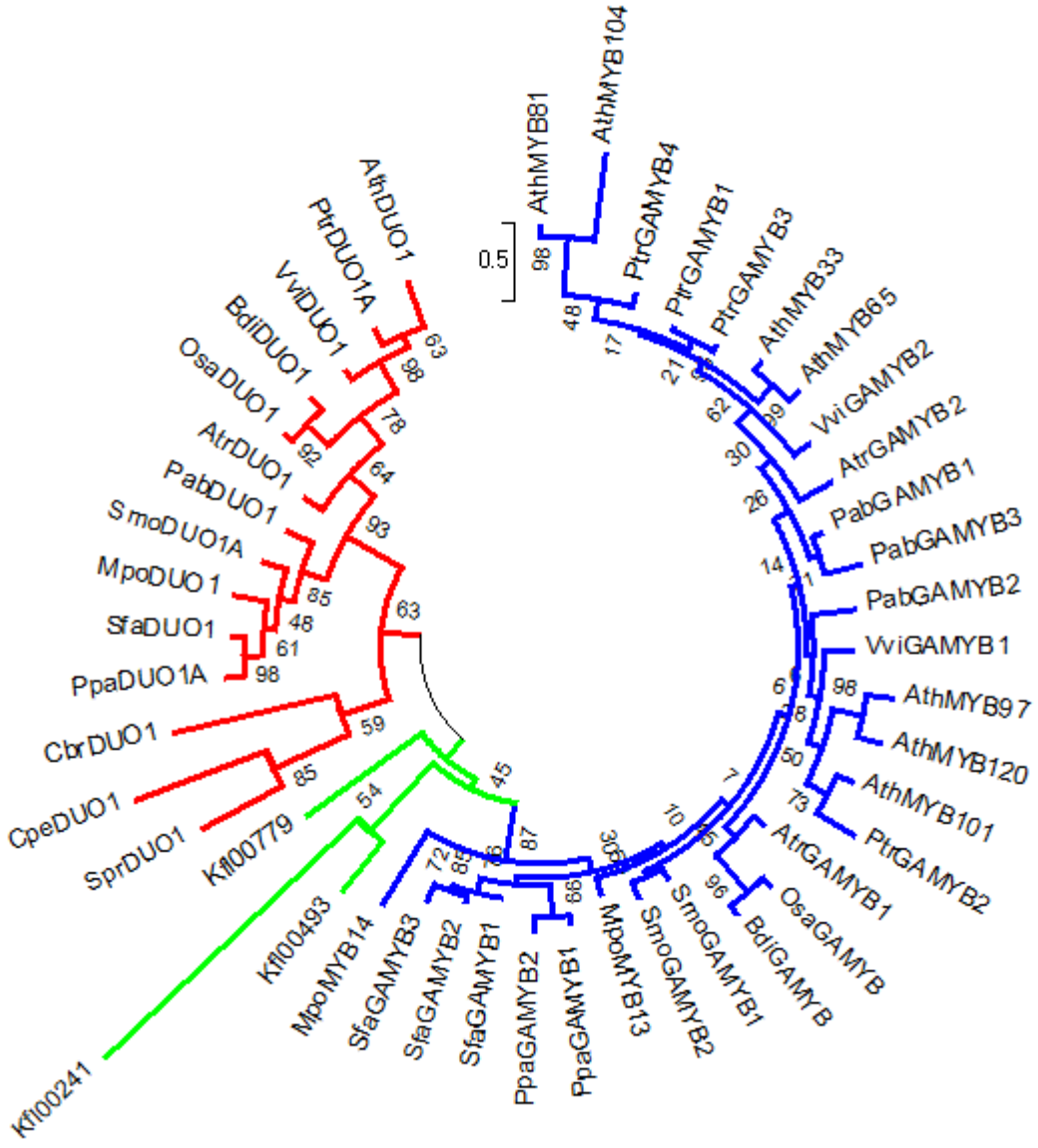


Figure 3. 9 Maximum likelihood tree for DUO1 and GAMYB phylogeny.
DUO1 (red) and GAMYB (blue) homologs in land plants form two distinctive clades (Bootstrap 93 and 87). Algal (green) homologs were selected from Charophyta, the known sister group to Embryophyta, to search for the early DUO1 homologs. Representative species were chosen from major clades, including monocot, eudicot, basal angiosperm, gymnosperm, lycophyte, bryophyte, Zygnematophyceae, Charophyceae, and Klebsormidiophyceae. Homologs from *Chara*, *Spirogyra*, and *Closterium* are grouped with the DUO1 clade, consistent with the presence of the K66 in their sequences. The closest MYB homologs from Klebsormidium, however, do not appear to belong to either the DUO1 or the GAMYB clade. The phylogenetic tree was constructed using maximum likelihood with Generalised Time Reversible (GTR) Model and bootstrap (1000) values are shown at the nodes.

3.5 Expression patterns of DUO1 in land plants

Expression patterns can be useful indicators of functional conservation. A change in the expression location would suggest a change in the protein function. The presence of DUO1 in *Arabidopsis thaliana* is strictly within the male germline cells. Some transcriptome studies suggest that this strict expression pattern is not only within Brassicaceae. In strawberry *Fragaria vesca*, DUO1 (gene12261) showed anther (late stages)-specific expression and the expression level was higher in isolated pollen (Hollender et al., 2014). In a grapevine microarray data analysis, hierarchical clustering (HCL) showed VvDUO1 expression was pollen-specific (Fasoli et al., 2012). In the Asterids family Solanaceae, DUO1 expression was only found in pollen in *Nicotiana tabacum* using microarray (Hafidh et al., 2012), and only in flowers for the two DUO1 homologs in *Solanum lycoperiscum* (Bostan and Chiusano, 2015).

In monocots, DUO1 was detected in *Zea mays* pollen (Table 3. 2) (Davidson et al., 2011), and this was independently confirmed by another RNA-seq experiment (Chettoor et al., 2014). For *Oryza sativa*, the OsDUO1 was found in the sperm cells (Anderson et al., 2013).

Table 3. 2 RNA-seq data from Davidson et al (2011) show that two copies of the maize DUO1 homologs have no pollen expression, indicating a loss-of-function.

	Gene	Chr	Pre-tassel	Post-tassel	Anther	Pollen
ZmDUO1A	GRMZM2G105137	chr2	9.8	7.0	19.1	181.5
ZmDUO1B	GRMZM2G046443	chr1	13.4	6.0	0.0	0.0
ZmDUO1C	GRMZM2G311059	chr10	10.0	0.0	0.0	0.0

The *Lilium longiflorum* generative cells EST research showed two possible AtDUO1 homologs (Okada et al., 2006), GC0489 and GC1562 with the similarity E-value $<1.6\times10^{-32}$ and 4.0×10^{-47}, respectively. In the *L. longiflorum* RNA-seq and cDNA library study (Lang et al., 2015), LlDUO1 (Unigene16620) was more than 12 times enriched in pollen than the rest of the tissues.

In fact, this expression specificity can be traced back long before the emergence of the angiosperms. For bryophytes, in *Physcomitrella patens*, the PpDUO1A (PP1S16_281V6.1) is male specific while the PpDUO1B (PP1S114_136V6.1) showed male preferential expression (Ortiz-Ramirez et al., 2016). In *Marchantia polymorpha*, MpDUO1 (Mapoly0019s0071.1) is also male specific (Higo et al., 2016). Algal DUO1

homologs expression provided by Dr Asuka Higo further suggested that DUO1 was linked to male gametophytes from the very beginning.

On the BAR (The Bio-Analytic Resource for Plant Biology) website (http://bar.utoronto.ca/), more DUO1 homologs expression patterns can be found. A selected few are presented in the Appendix.

3.6 Diploidisation of the DUO1 clades

Not only the expression patterns are conserved, the copy numbers of DUO1 also tend to remain at a low number (Figure 3. 10), in contrast to the GAMYB family (Figure 3. 9). If it is the case, the strict role of DUO1 within the male germline may have prevent it from finding other functions in other tissues, which leads to its diploidisation.

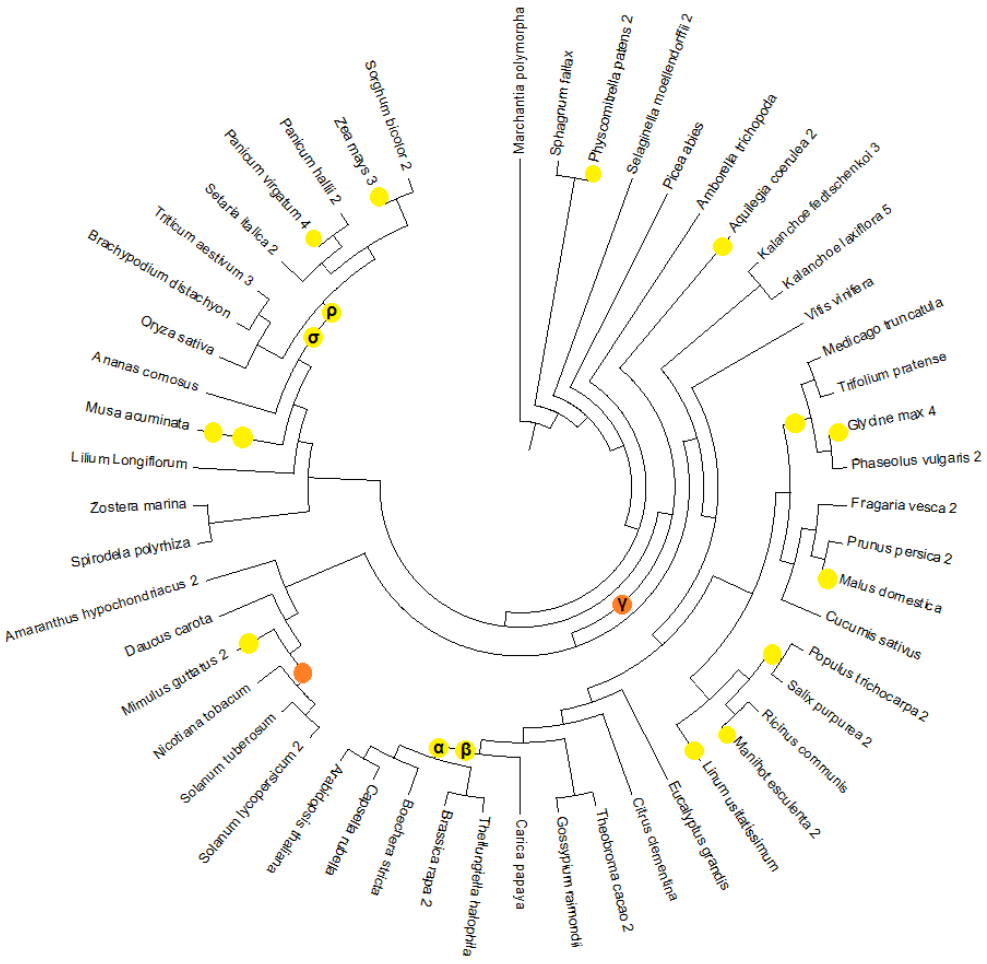


Figure 3. 10 DUO1 homologs are found in all different land plant species.
The cladogram is built based on the APG IV classification (Byng et al., 2016). The numbers after species names indicate those with multiply copies and their copy numbers. Yellow and orange circles indicate known whole genome duplication and triplication events respectively (see discussion in Chapter 1).

As discussed in Chapter 1, multiple whole genome duplication (WGD) events and, much rarer, whole genome triplication events have happened during the history of many clades in angiosperms. For example, two WGD have happened recently in the Brassicaceae, but instead of four copies of DUO1, most species examined only have a single copy except for *Brassica rapa*. The case of *B. rapa* is more complicated as its genome is triplicated (Wang et al., 2011).

To confirm this idea, the monocot clade Poales which experienced two recent ancestral duplication events, was chosen as an example. The positive selection test was to see if all the *DUO1* copies had a regular non-synonymous/synonymous substation ratio (ω). An ω value above 1 would indicate a higher mutation rate, which could be contributed through either a positive selection or a relaxed selection. The test was done by the Clade model and Branch-Site model using the PAML 4 package (Yang and dos Reis, 2011, Yang, 2007).

The results are presented in Figure 3. 11. There were two known ancestral duplication events σ and ρ in the monocot clade (Tang et al., 2010, Muhlhausen and Kollmar, 2013). Interestingly, one could only observe the effect of one event through the collection of *DUO1* sequences. It is possible that the extra copy from event σ was completely lost since one copy of DUO1 is functionally sufficient, and the same rapid diploidisation can also explain the single copy in *Oryza sativa* and *Brachypodium distachyon*. For clade 18-29, relaxation happened right after the duplication, suggesting a loss of function. This is consistent with the mutated C-terminus regions of PviDUO1C/D and the lack of ZmaDUO1B in pollen (Table 3. 2). On the other hand, in the clade 18-19, one extra copy was released of selection pressure after the duplication events in both *Panicum virgatum* (b) (Missaoui et al., 2005) and *Zea mays* (a) (Gaut and Doebley, 1997). Similarly ZmaDUO1C was also absent in the pollen transcriptome data. These examples show that DUO1 has a strong tendency of diploidisation. However, relaxation were not detected in any of the three copies from *Triticum aestivum*, likely due to the fact that the hybridisation event happened less than half a million years ago (Marcussen et al., 2014, International Wheat Genome Sequencing, 2014), and the domestication started only around 10,000 years ago (Feldman and Kislev, 2007), too recent to allow many mutations to accumulate. Surprisingly, a positive selection was detected for the rice DUO1. Human cultivation could be behind this as reproduction advantage has always been rigorously selected by farmers.

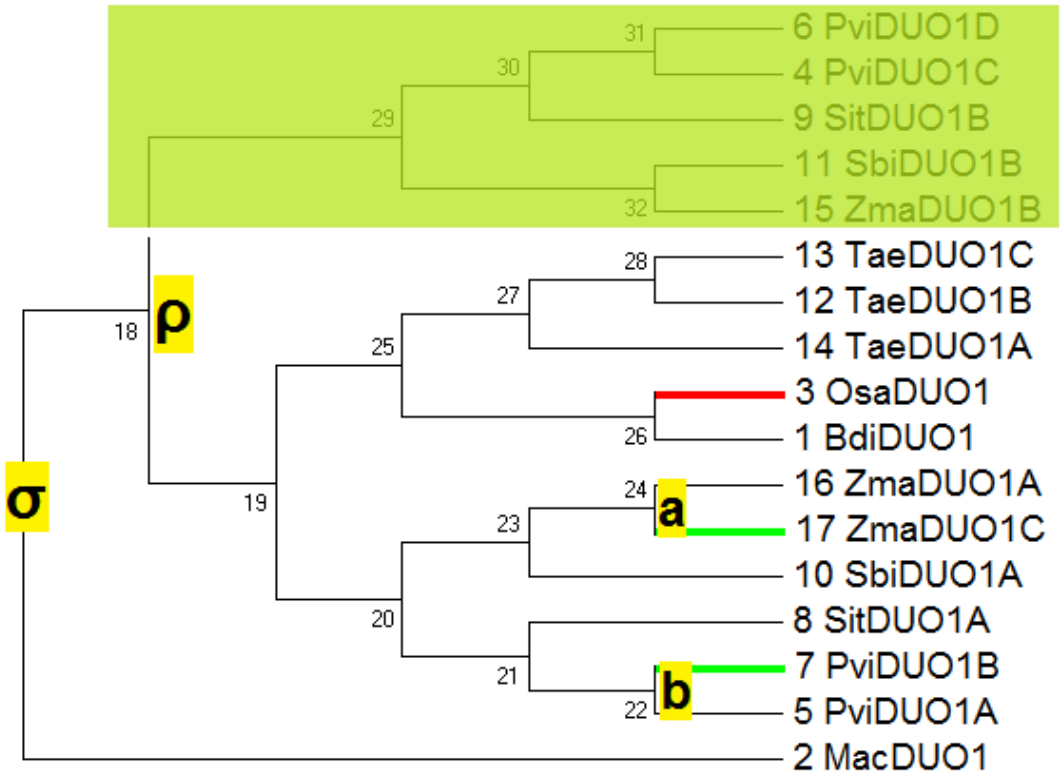


Figure 3. 11 Positive selection and relaxation on different branches of Poales.
The green box indicates elevated ω value in the branch in the Clade model test. Individual branches were then tested with the Branch-Site model. The green branches indicate relaxation, the red branch indicates positive selection detected (details see Appendix). The yellow boxes mark the known whole genome duplication events. Bdi=*Brachypodium distachyon*, Mac=*Musa acuminata*, Osa=*Oryza sativa*, Pvi=*Panicum virgatum*, Sit=*Setaria italica*, Sbi=*Sorghum bicolor*, Tae= *Triticum aestivum*, Zma= *Zea mays*.

However, the power of selection pressure, or rather the lack of it, does not explain why the diploidisation of DUO1 was so rapid and happened in so many independent clades.

In comparison, the GAMYB family seems to have proliferated into a huge clade after the WGD events in the angiosperms. That is not to say the GAMYB did not experience any loss-of-function events within the clades. For example, there are two homologs in the basal angiosperm *A. trichopoda*, but after two WGD in the Brassicaceae there are only seven homologs instead of eight in *A. thaliana*. Many GAMYB homologs were being lost as a result of the selection pressure being relaxed, and it is important to have a look at the speed of the loss to better understand how fast extra DUO1 copies are purified.

As shown in Chapter 1, for the seven homologs in *A. thaliana*, functional redundancy among MYB97, MYB101, and MYB120 were documented (Allen et al., 2007, Leydon et al., 2013, Liang et al., 2013), one copy is most likely lost in this sub-group. Separately, MYB33 and MYB65 are also redundant to each other (Millar and Gubler, 2005, Allen et al., 2007). So far, it is not clear whether redundancy exists between MYB81 and MYB104. In fact, evidence based on sequence analysis and expression patterns suggested that MYB104 might be drifting away.

In the genus *Arabidopsis*, no homolog for AtMYB104 was found in *A. lyrata*, while the homolog in *A. halleri* showed many SNPs within the MYB domain indicating an elevated ω value, and a two-amino-acid deletion in the first α-helix of the R2 repeat which would cause instability of the MYB hydrophobic core. In the genus *Capsella*, both *C. rubella* and *C. grandiflora* have only one homolog in this clade, a similar situation as *A. lyrata* (Figure 3. 12).

More information can be drawn from the microarray data (Twell, unpublished) of the male reproduction tissues (Table 3. 3). For all members of the *A. thaliana* GAMYB family, MYB33/65 have an early expressing profile, while MYB97 and 101 are highly expressed at a later stage. These expression patterns conform to the functional redundancies within the sub-clades. In contrast, MYB81 is expressed early in the pollen development, but MYB104 is not properly expressed at any stage.

Table 3. 3 Microarray data for the *A. thaliana* GAMYB proteins shows the lack of MYB104 expression, a sign of possible degeneration.

	ATH1	ATH1	ATH1	ATH1	ATH1	ATH1	ATH1	ATH1
GENE	RT	SED	LF	UNM	BCP	TCP	MPG	SC
MYB81	123.99558	127.61022	176.79341	348.36	316.71	232.9	348.73	131.73
MYB104	189.63918	200.17492	171.08441	195.48	172.31	269.97	328.09	309.79
MYB33	148.71945	332.40029	181.63719	211.39	289.5	339.48	298	54.8
MYB65	352.39173	526.36104	322.99019	653.14	710.33	687.81	852.02	433.49
MYB101	121.92441	256.78988	128.69023	269.88	320.96	2563.78	5505.05	213.03
MYB120	147.08601	169.52675	140.66951	74.31	65.45	115.76	212.85	112.43
MYB97	76.187873	95.122179	80.996561	100.5	113.35	532.66	1415.11	123.06

RT = root; SED = seedling; LF = leaf; UNM = uninucleate microspore; BCP = bicellular pollen; TCP = tricellular pollen; MPG = mature pollen grain; SC = sperm cell.

A similar fate was found for the *Arabidopsis* MYB120 homologs, agreeing with previous studies that showed it was not a transcription activator (see Chapter 1).

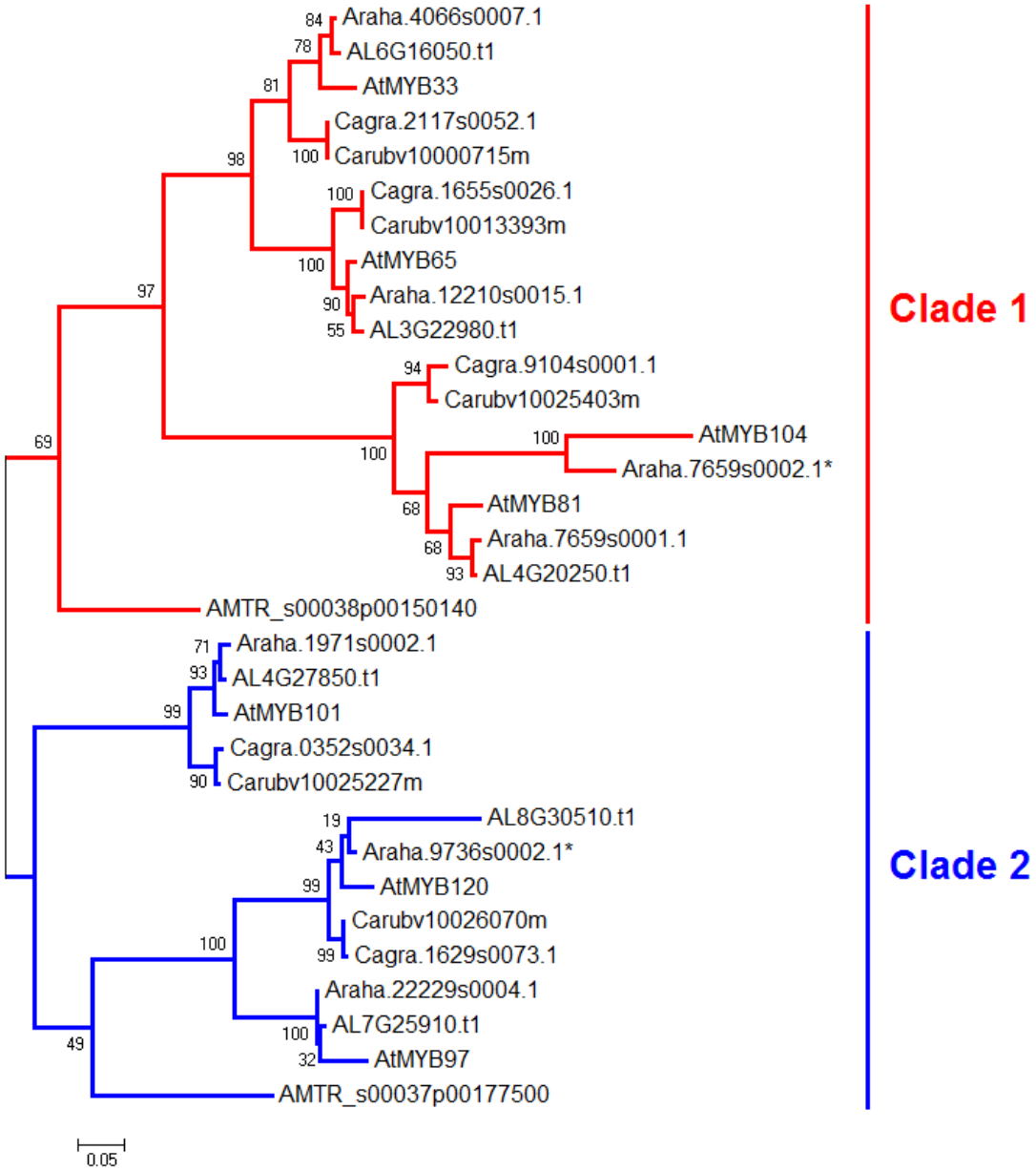


Figure 3. 12 The proliferation of GAMYB-like family in Brassicaceae.
In basal Angiosperm *Amborella trichopoda*, only two homologs of the GAMYB proteins exist. They became two clades in Brassicaceae after two whole genome duplication events. For most GAMYB family members in *Arabidopsis thaliana*, there is a homolog in close-related species. Notably, AtMYB104 has no homolog in *Arabidopsis lyrata*, and only one homolog for the pair of AtMYB81/104 in the genus *Capsella*. This suggests that MYB104 homologs are degenerating in Brassicaceae species. The phylogenetic tree is constructed with maximum likelihood using Tamura-Nei Model and 1000 bootstraps (Tamura and Nei, 1993). All nucleotide sequences were from phytozome v12.0 (phytozome.jgi.doe.gov), including *Amborella trichopoda*, *Arabidopsis halleri*, *A. lyrata*, *A. thaliana*, *Capsella grandiflora*, and *C. rubella*. * indicate predicted loss-of-function mutations from sequence analysis, due to possible degeneration of the genes.

3.7 Discussion

DUO1 first emerged within the fresh water green algae Charophyta. Dr Asuka Higo hypothesized the separation of DUO1 and the GAMYB family after a potential duplication. It is possible that the function of DUO1, though evolved in the past 500 million years, has been conserved to control the male germline cell fate. The control over cell cycle, on the other hand, is more likely to be a recent innovation accompanies the rise of the angiosperms, as the network has changed since bryophyte (Dr Asuka Higo, unpublished).

This hypothesis is in agreement with the structural conservation of the DUO1 proteins. The MYB domain, which is responsible for the target recognition has changed little since the Charophyta and the target binding sequences have remained the same since the bryophytes. In the next chapter, the position weight matrices of the top-scoring 8-mer DNA sequences is present done by my collaborator, bound by different MYB domains on a protein binding DNA microarray. These matrices have confirmed the binding sequences of both AtDUO1 and MpDUO1 are TAACCGT/CT/C, the same sequence as discussed previously in this chapter. However, the C-terminus that contains a possible activation domain has evolved both in sequence and in function, which will be discussed in Chapter 6.

The concept of sex-biased gene expression has been well established through studies in animals (Ellegren and Parsch, 2007, Parsch and Ellegren, 2013). It was also found in brown algae (Martins et al., 2013, Ahmed et al., 2014, Lipinska et al., 2015, Lipinska et al., 2016), even for isogametes (Lipinska et al., 2013). These reproductive genes have also been reported to evolve at a high rate (Swanson and Vacquier, 2002, Wilburn and Swanson, 2015). For example, in *Chlamydomonas reinhardtii* the MID (MINUS DOMINANCE) decides the mating types and FUS1 is essential for the fusion of the + and – type cells, yet no homolog of *FUS1* and only one homolog of *MID* was found in 12 other species of *Chlamydomonas* (Ferris et al., 1996, Ferris et al., 1997). In most cases, the number of male-biased genes is higher than the one of female-biased genes (Ellegren and Parsch, 2007), following the "Bateman's Principle" (Bateman, 1948), with some exceptions involving complex mating strategies like in human (Brown et al., 2009).

Klebsormidium reproduce using zoospores (Marchant et al., 1973, Rogers et al., 1980, Rindi et al., 2008). The sexual reproduction in this group is still unknown. Zygnematales

like *Spirogyra* achieve sexual reproduction through conjugation (McCourt et al., 2000, Gontcharov et al., 2003), and Charales through heterogametes (McCourt et al., 2004, Kapraun, 2007). Based on the analysis of the homologous sequences presented earlier in this chapter, it is likely the earliest DUO1 homologs evolved after the separation of Klebsormidiophyceae and Charophyceae. Thus, the emergence of DUO1 was correlated with the beginning of sexual reproduction in Streptophyta. DUO1, along with the GAMYB homologs, could be the genes linked to heterogamy or sexual reproduction in this clade.

However, while *GAMYB* family proliferated over time, e.g. 7 homologs in *Arabidopsis thaliana*, which fits the profile described above, *DUO1* homologs remained a low copy number within most species. After any polyploidy events, *DUO1* has a strong tendency of rapid diploidisation as demonstrated in the example of the monocot Poales. This lack of neo-functionalization suggests that the role of DUO1 might be much more fundamental and specific in the evolution of the male germline development. This could mean that a high dosage of DUO1 is potentially deleterious, which is supported by the ectopic expression experiments that showed activation of germ cell-specific gene expression in many non-germline cells (Brownfield et al., 2009a).

Chapter 4 Analysis of DUO1 Functional Conservation in Bryophytes

Abstract

DUO1 homologs are known to regulate male germline development in some angiosperms. Two independent pathways for cell cycle progression and differentiation in the male germline cells require the presence of DUO1 in *Arabidopsis thaliana*. Inside the angiosperm group, different DUO1 homologs are able to complement the phenotype of the *A. thaliana duo1* mutant *in planta*, and activate target promoters *in vivo*. For pre-angiosperms, the double fertilization process has never been observed, therefore it is unclear whether DUO1 homologs in these plants regulate cell division. The DUO1 homologs analyzed in these pre-angiosperms have the same conserved MYB domain but the C-terminal end contain less acidic residues. DUO1 homologs from the bryophytes like *Physcomitrella patens* and *Marchantia polymorpha* showed the ability to recognize and bind to the same target promoter DNA sequences in *A. thaliana*, but could not activate them efficiently *in vivo* or rescue the cell differentiation defect phenotype in *A. thaliana*.

4.1 Introduction

As described in Section 1.2, the function of DUO1 homologs as a regulator for cell division and cell differentiation for the male germline development process can be easily confirmed by complementing the *duo1* phenotypes. Dr Ugur Sari has previously performed functional complementation experiments, and found that both the rice and one of the tomato (A) *DUO1* homologues were able to rescue both failure of germ cell division and differentiation in *A. thaliana duo1-1* pollen (Sari, 2015).

This *in planta* complementation is a good tool for evaluating certain molecules function capability in a native context. Although it is not possible to quantify the transcriptional regulation once the complementation reaches full rescue percentage, it measures the overall biological functional activity by testing if the *duo1* mutant with the transgene is fertile. Note that the signal quantification is still not achievable directly through the presence of the GFP marker, as there is little one can do to distinguish the two classes of "*DUO1*, RFP+ (transgene positive) GFP+ (transactivation positive)" and "*duo1*, RFP+ GFP+" pollen at the F1 generation.

Another approach, the dual-luciferase transient assays were used in combination with the *in planta* complementation. For instance, as a much more time efficient method, molecules were often tested in the transient assays to see if they had any activation potential before creating any transgenic lines. For the same reason, it was also a preferred choice when a large number of constructs needed to be tested. In certain cases, artificial molecules were created to test if the alterations have any effect on the transactivation ability, so the quantitative measurement was more important than the native context to test the differences between the activities of the molecules.

In Chapter 3, it was believed that DUO1 homologs had evolved in the green algae and fixed into its earliest recognisable form in the bryophytes. Many more functional test data have been generated for the angiosperm DUO1 homologs, yet without being summarised for an overall view previously. However, a main function of the DUO1 homologs in the angiosperms is to regulate the cell cycle progression during the pollen mitosis II (see Chapter 1). Pre-angiosperms do not have any pollen grains and thus the cell cycle progression of the germline cells are likely to have some differences.

In this chapter, previously unpublished data were collected and summarised to gain a broad understanding for the role of DUO1 within the angiosperm clade. Analysis of the DUO1 functions were also done in the bryophytes, representing the early land plants which the gametophyte stage dominates the life cycle, and the swimming sperms develop into maturity without the need for a series of cell divisions.

4.2 Summary of functional conservation of DUO1 in angiosperms

To standardise the *in planta* complementation rescue efficiencies, a chart with both the tricellular pollen (TCP) grain percentage and the *HTR10* promoter activation percentage is presented in Figure 4. 1. This example is a reinterpretation using the data from (Vesty, 2010). The percentage of the TCP and GFP-expressing pollen grains were counted and they represent the cell division rescue and the cell differentiation rescue, respectively. As expected, the negative control *duo1-1*$^{+/-}$ line had about 50 % TCP, GFP+ and 50 % BCP, GFP-. For the single insertion lines, half of the BCP, GFP- (25 %) type harboured the transgenes. If the transgene had full rescue abilities (e.g. AtDUO1), the TCP, GFP+ would become 50 % + 25 % = 75 %.

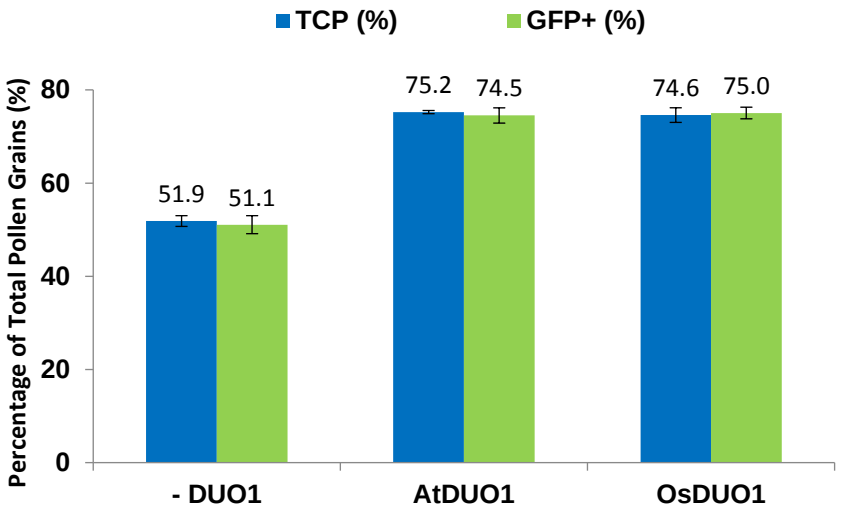


Figure 4. 1 OsDUO1 functional test results reinterpreted to show the rescue of germ cell division and differentiation defects in *Arabidopsis thaliana*.
Original data were collected from (Vesty, 2010) and reorganised for presenting in the fashion of this chart. The numbers show frequencies of tricellular pollen (TCP) and promHTR10:H2B-GFP-expressing (GFP+) pollen in heterozygous *duo1-1* plants, harbouring promDUO1:AtDUO1-mCherry (AtDUO1), promDUO1:OsDUO1-mCherry (OsDUO1), or without any construct (- DUO1). Error bars represent the SE. n ≥ 2 lines.

A series of DUO1 homologs (also see Chapter 5) were tested using the dual-luciferase transient assays. All the angiosperm DUO1 homologs tested, as in Figure 4. 2, showed abilities to activate *Arabidopsis thaliana* (Rosids) DUO1 target promoters, including *Solanum lycopersicum* (Asterid), *Oryza sativa* and *Zea maize* (commelinids), *Lilium longiflorum* (Lilioid), and *Amborella trichopoda* (basal angiosperm).

Combined with the functional complementation results, it is reasonable to conclude that the core functions of DUO1 are largely conserved within the angiosperms. Interestingly, the DUO1 of *A. trichopoda* failed to express *in planta* (with C-terminal mCherry fusion) or *in vivo* (with C-terminal GFP fusion), similar to the case of SlDUO1B (Chapter 1).

However, this leads to the questions regarding the functions of the pre-angiosperms DUO1 homologs. First, double fertilisation has never been observed outside the angiosperm clade, i.e. the cell cycle progression may not be required. Second, the emergence of DUO1 coincided with the establishment of sexual reproduction with the Streptophyta clade (Chapter 3). Finally, the sperm cells retained the ability to swim for a very long period, is this dramatic change reflected on DUO1 on any level?

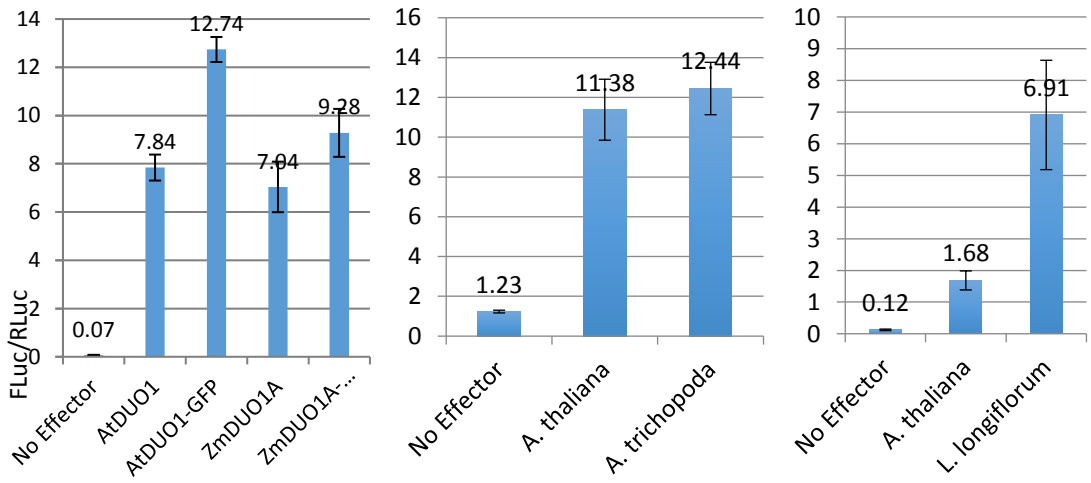


Figure 4. 2 Further tests of angiosperm DUO1 homologs activating *Arabidopsis thaliana* promoters in tobacco leaves.
DUO1 homologs from *Zea maize* (A), *Amborella trichopoda*, and *Lilium longiflorum* were tested separately on a known DUO1 direct target *HTR10* promoter. All of them demonstrated various abilities of activation. The ZmDUO1A was also tested with a GFP tagged at the C-terminal, and was selected after being predicted as the only functional copy of the three maize homologs in Chapter 3. Raw data collected by Mowaninuolajesu Ojo (*Z. maize*), James Walker (*A. trichopoda*), and Mikhaela Neequaye (*L. longiflorum*).

4.3 *In vivo* test of *Physcomitrella patens* DUO1 homologs

To understand more about the DUO1 homologs from pre-angiosperms, the first model bryophyte chosen was the moss *Physcomitrella patens*. There are two homologs in *P. patens*, likely due to a recent whole genome duplication (WGD) event.

The cDNA of both genes were obtain by Dr Sari and contain slight variations compared to the sequence predictions available (Sari, 2015). However, note that the nomenclature used in his thesis was "Pp1S114_136V6 and Pp1S16_281V6 were named as *PpDUO1-A* and *PpDUO1-B*". This is the exact opposite of the nomenclature used in this thesis (namely *PpDUO1A*: Pp1S16_281V6 and *PpDUO1B*: Pp1S114_136V6). This was to name the genes according to the GO numerical order, strictly for the purpose of easy communication with other academics involved or might be involved in relevant works. All names used in the thesis were based on this nomenclature.

The complementation for both genes were carried out by Dr Sari. However, PpDUO1B (Pp1S114_136V6, named A in his thesis) failed to express the mCherry signal, and PpDUO1A (Pp1S16_281V6, named B in the thesis) did not show any rescue for the cell division defect (~50 % TCP) when expressed. It was not clear what had caused the failure of expression of the PpDUO1B.

To circumvent this problem, two T-DNA vector based constructs pB2GW7-PpDUO1A and pB2GW7-PpDUO1B were made. Both constructs contained a CaMV 35S promoter to drive effector expression, which was the same promoter used to drive the *Renilla* luciferase expression in the internal control plasmid (pB2GW7-RenLuc). This resolves the potential promoter expression difference by simply normalizing the results. The constructs were then tested in transient assays using tobacco leaves (Figure 4. 3). The transient assays were independently repeated by Mikhaela Neequaye and showed similar results (see Appendix).

The *Arabidopsis thaliana* DUO1 served as a positive control, which gave a standard level of activation above the no effector negative control. Neither constructs of *P. patens* DUO1 demonstrated any ability to increase the target *HTR10* promoter activity after being introduced into the infiltration mixtures.

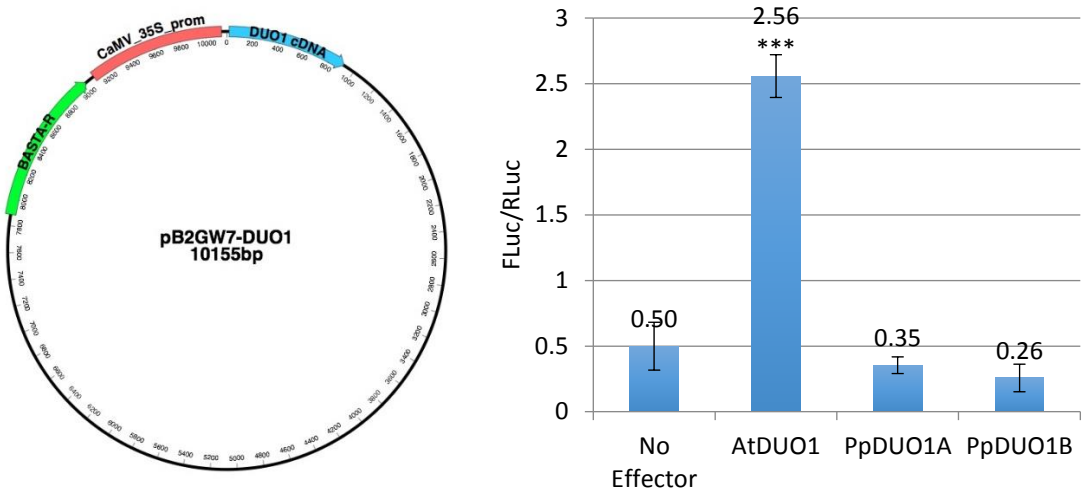


Figure 4. 3 The two constructs from *P. patens* failed to activate AtDUO1 target *HTR10* promoter *in vivo*.
All three homologs were transformed into a pB2GW7 vector. AtDUO1 served as a positive control. Neither moss DUO1 effectors showed a Fluc/RLuc ratios level significant deviated from the negative control with no effector. *** $p<0.01$.

These results suggested that neither the PpDUO1A nor PpDUO1B have the functional activity which is conserved within the angiosperm. However, as described in Chapter 3, the MYB domains of DUO1 homologs from *A. thaliana* and *P. patens* (similar to *M. polymorpha*) share more than 70 % in identity and 80 % in similarity. The protein models of the two can also be overlapped well (Figure 4. 4 A). Thus, it was suspected that the MYB domain of the PpDUO1 homologs could still recognize and bind to the same target sequence as AtDUO1.

To demonstrate if the PpDUO1 MYB domain can bind to the same target, an artificial chimeric molecule was made (Figure 4. 4 B), with PpDUO1B N-terminal region stops after the MYB domain, and the C-terminal region of the AtDUO1, referred to as "PpChimera" in the following context. This was achieved by amplifying the two relevant fragments through PCR, using primers designed with overlapping overhangs (see supplementary). Then using both fragments as templates in a single reaction, amplify with the attB1 and attB2 adapter primers. After the optimization for the thermocycler programming, the conditions were set as normal except for the first ten repeating cycles the annealing temperature at 45 °C for better hybridisation between the templates, and the rest cycles return to normal at 55 °C annealing temperature.

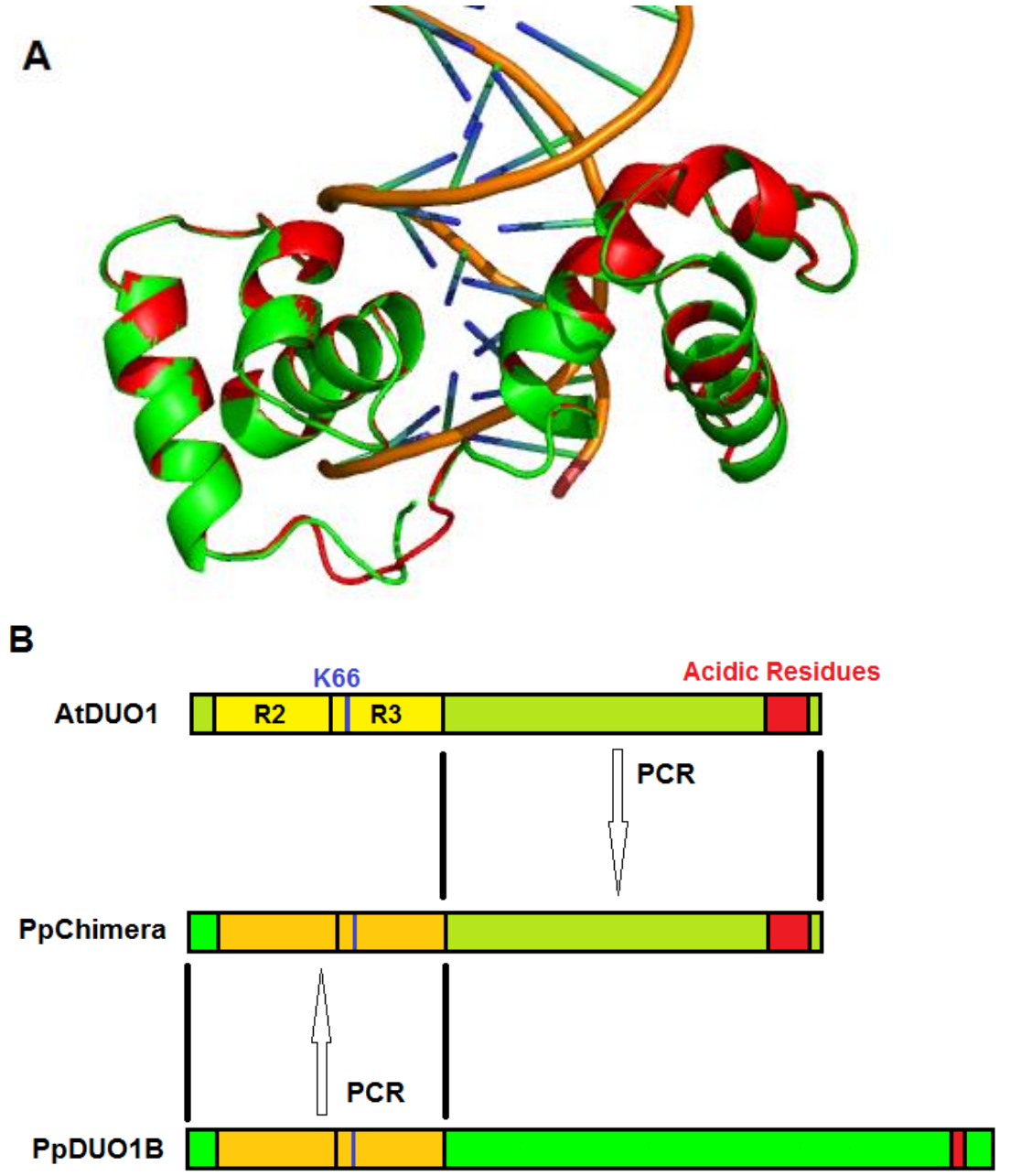


Figure 4. 4 PpDUO1B N-terminal region (MYB domain) was used to fuse with AtDUO1 C-terminal region to create a chimeric molecule.
A. Protein models of PpDUO1B and AtDUO1 MYB domains overlaps well. Red = AtDUO1 MYB; Green = PpDUO1B MYB. B. The chimeric molecule was created using PCR, the alignment shows where the conjunction is right after the MYB domain. Yellow = MYB domains; Red = C-terminal motifs.

This chimeric molecule was then tested in the transient assays with the *HTR10* promoter. Although this approach could not give back the exact binding sequence of the PpDUO1B

MYB domain, it could provide information of its *in vivo* ability to activate the same target as AtDUO1, of which the binding sequence is known (Borg et al., 2011). The results are shown in Figure 4. 5.

The PpChimera construct was able to raise the activation level of the target *HTR10* promoter to ~50 % of the AtDUO1 (positive control) level, in comparison to PpDUO1 homologs with no evidence of activation. When the concentration of the PpChimera construct was increased to four times of the standard amount by using a higher concentration of cells, the activation level was increased further above the positive control AtDUO1. Although the response was not linear, it demonstrated that the PpChimera construct was responsible for the increase of the activation level.

This result agreed with the aforementioned hypothesis that the PpDUO1B MYB domain can recognize and bind to the same target promoter sequence as AtDUO1, the difference in the activation abilities between AtDUO1 and PpDUO1 homologs were mainly due to the C-terminal region of the molecules. This is further demonstrated and discussed later.

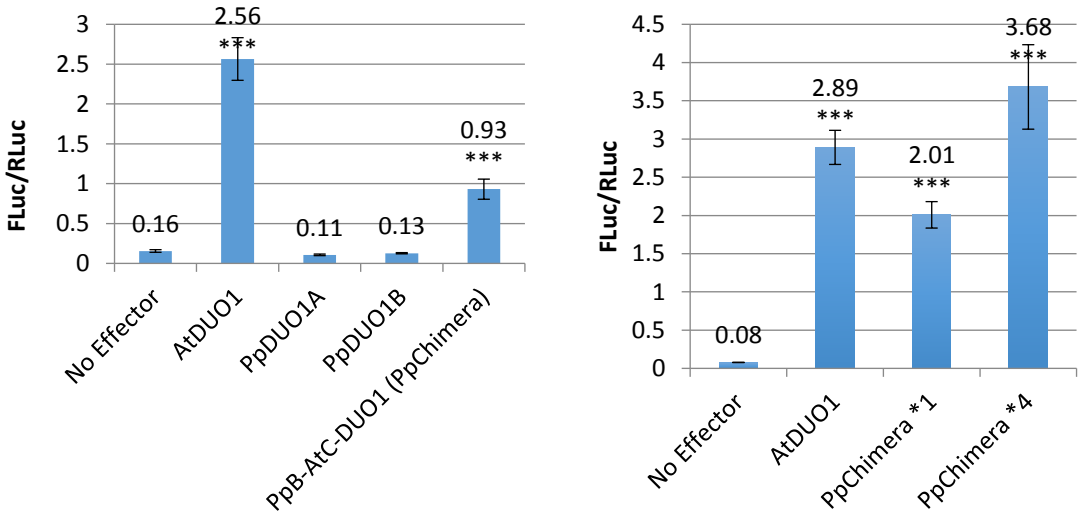


Figure 4. 5 PpDUO1B N-terminal region (MYB domain) fused with AtDUO1 C-terminal region creates a chimeric molecule that has *in vivo* activity.
The PpChimera molecule was able to activate the target *HTR10* promoter to ~ 50 % of the AtDUO1 level, while PpDUO1B with the same N-terminal region failed to activate this target. The increase of concentration of the PpChimera constructs also led to an increased activation level of the target promoter, indicating the activation increase was the direct result of the presence of the PpChimera. This demonstrated that PpChimera is an activator and its MYB domain (PpDUO1B MYB) is able to recognize and bind to the same target sequence as AtDUO1 MYB domain. *** $p<0.01$; *4, four times effector concentration.

4.4 Knockout test of the *Physcomitrella patens* DUO1 homologs

Even though PpDUO1 homologs failed to activate AtDUO1 targets *in vivo* or *in planta*, they are still very likely regulating the male germline development in *P. patens*, interacting with a different network of genes.

Genes of *P. patens* can be accurately targeted through homologous recombination (HR) as discussed in Chapter 1 and 2. The moss transformation vectors used in this work were the same as and received from Schaefer, 2010 (Figure 4. 6). They contain ~1000 bp of 5' and 3' genomic fragments of the native target genes (PpDUO1 homologs), flanking a 35S:nptII resistence-mediating cassette (Lindner et al., 2014). The plasmids pBNRf and pBHRf have a 35S:neoR and a 35S:hygroR cassette respectively, cloned between two LoxP sites in a pMCS5 derivative backbone (Schaefer et al., 2010).

The targeting arms of the two *PpDUO1* homologs were designed as below and then amplified using PCR.

For pBNRf, restriction sites BamHI and XhoI were designed for the 5' targeting arm, NotI and PacI for the 3' targeting arm. This was used to target Pp1s16_281V6.1 (*PpDUO1A*). *PpDUO1A* 5' targeting arm starts from 915 bp upstream of the start codon ATG and ends at ATG (exclude, same as for the PpDUO1B). *PpDUO1A* 3' targeting arm starts 702 bp upstream of the stop codon and ends at 192 bp downstream of the stop codon.

For pBHRf, restriction sites HindIII and XhoI were designed for the 5' targeting arm, NotI and PacI for the 3' targeting arm. This was used to target Pp1s114_136V6.1 (*PpDUO1B*). *PpDUO1B* 5' targeting arm starts from 1343 bp upstream of the ATG and ends at ATG. *PpDUO1B* 3' targeting arm starts 300 bp upstream of the stop codon and ends at 1123 bp downstream of the stop codon.

In both cases, the targeting arms would ensure the complete disruption of the MYB domain, which was shown to completely disrupt the DUO1 functions in Borg, 2011.

The plasmids were cut using BamHI and HindIII prior to the transformation, as transformations using circular plasmids have a much higher frequencies of random transgene integration, in comparison linear plasmids can ensure more gene-targeted insertions (Smidkova et al., 2010).

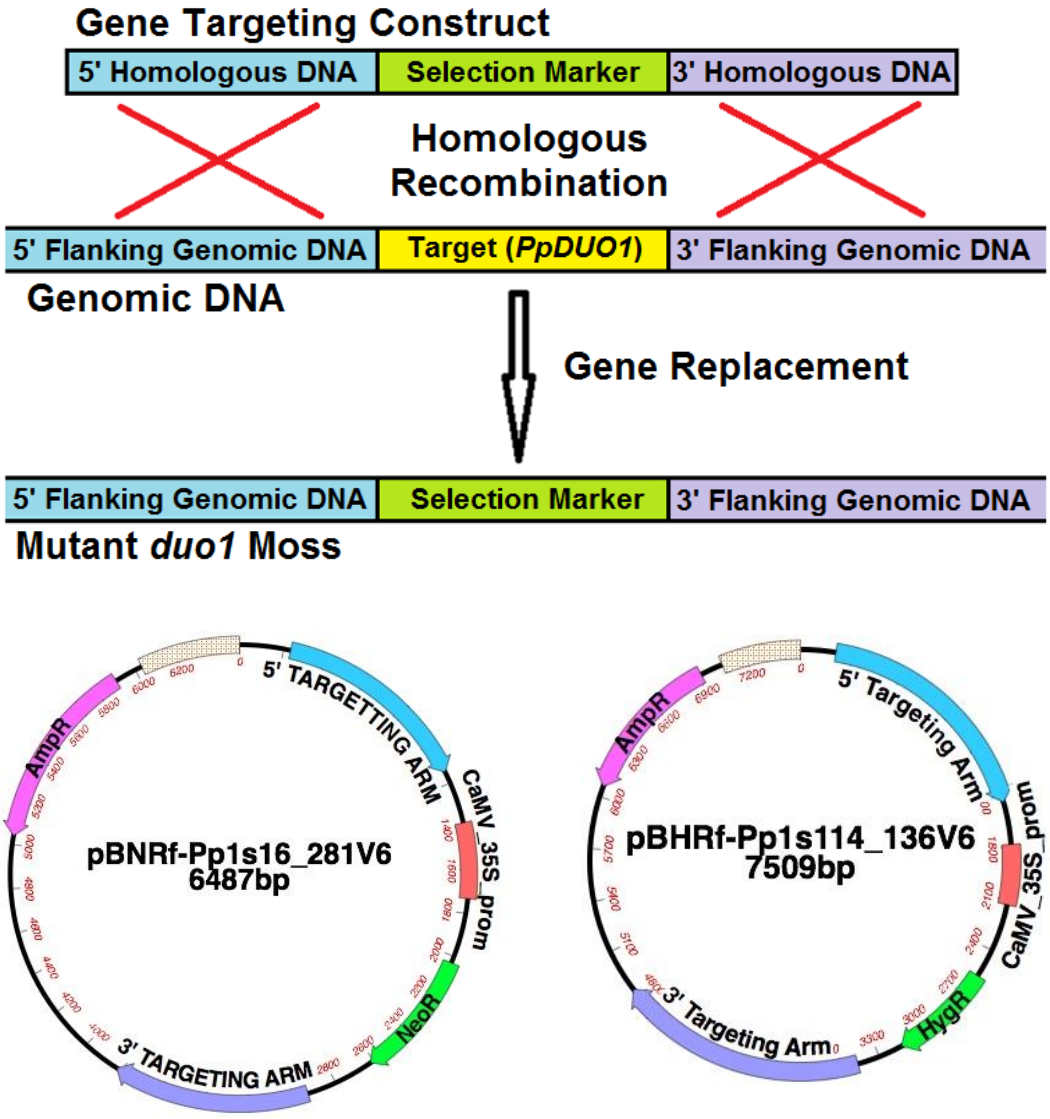


Figure 4. 6 Gene targeting in *Physcomitrella patens* using homologous recombination. Homologous targeting arms (blue and purple) are based on the genomic sequences surrounding the target genes. The two constructs targeting the two *PpDUO1* homologs both carry targeting arms ~ 1000 bp. After transformation, the selection marker (green) in the successfully knocked out lines would express the relevant antibiotics.

PpDUO1A was successfully knocked out. This was verified using PCR by Dr Lindner (see Appendix). However, the knockout lines did not show any obvious phenotype (Figure 4. 7). *PpDUO1B* did not yield a knockout line due to contamination during the selection process.

In Chapter 3, it was shown in the microarray data that the *PpDUO1A* was expressed exclusively in the male tissues, while the *PpDUO1B* signal was also detected in the female tissues. However, the *PpDUO1A* knockout line was fertile and able to produce spores. This indicated that PpDUO1A is not essential for antherozoid development,

which may imply that PpDUO1A and PpDUOB are functionally redundant, despite the expression pattern difference.

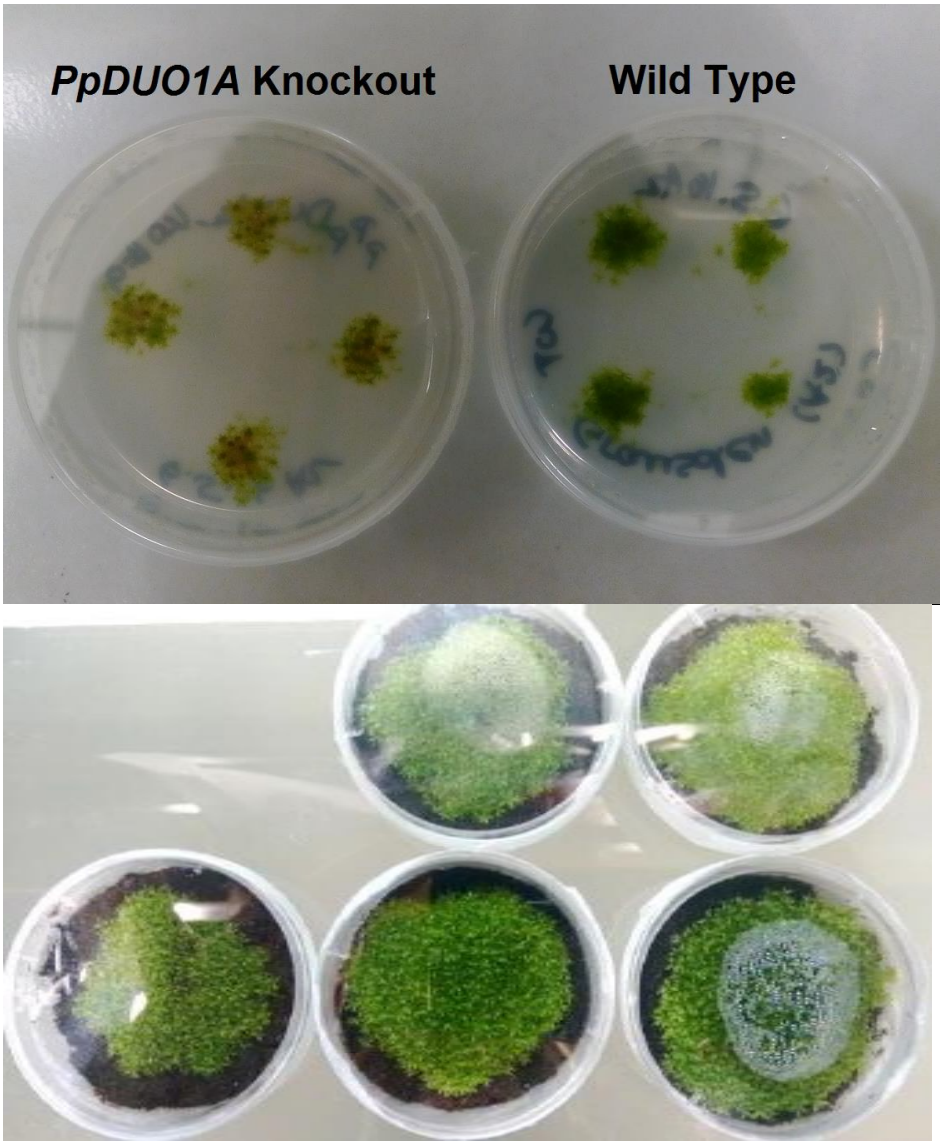


Figure 4. 7 The single knockout line for the *PpDUO1A* in *Physcomitrella patens* did not display any obvious phenotype and was fertile.
However, the understanding about the fertility is very limited, as the only knowledge gained is that the sperms of this knockout line was able to produce spores. *PpDUO1A* is male specific, instead of male preferential like *PpDUO1B* (Chapter 3), which indicates the redundancy between the homologs. The photos show one mutant line compared to a wild type (up) and other mutant F1 lines (down, picture provided by Dr Lindner).

4.5 *In vivo* test of *Marchantia polymorpha* DUO1 homolog

In recent years, the rise of CRISPR technology provided the world with new opportunities to explore. *Marchantia polymorpha*, for instance, has become a new model organism for bryophytes with only a single *DUO1* homolog.

The focus on the bryophyte for this work was also shifted to *M. polymorpha* in collaboration with Dr Asuka Higo from Prof Takashi Araki's group of Kyoto University and Dr Tomokazu Kawashima and Dr Michael Borg from Prof Frederic Berger's group of the Gregor Mendel Institute (GMI) in Austria. The following content has been submitted but not published.

In *M. polymorpha*, one *DUO1* homolog (Mapoly0019s0071.1) and two *GAMYB* homologs (initially labelled *MpMYB13* and *MpMYB14*, the nomenclature adopted in this thesis, available on Phytozome as Mapoly1089s0002.1 and Mapoly0023s0101.1) have been identified (Chapter 3).

The binding ability of the MpDUO1 MYB domain was first tested in the same way as the PpDUO1 MYB domain. A chimeric molecule using the N-terminal region including the MYB domain from MpDUO1 fused with the rest C-terminal region from AtDUO1, MpChimera, was created by PCR.

The GAMYB homologs were predicted to have a different targeting sequence, due to the differences in certain residues in the MYB domain. Therefore, a similar chimeric molecule, GAChimera, was made using the N-terminal region of the MpMYB13. This molecule was tested alongside with MpChimera. Diagrams of all relevant molecules are shown in Figure 4. 8.

MpDUO1, in contrast to the PpDUO1 homologs, was able to activate the *A. thaliana HTR10* promoter, although it was an even weaker activator compared to OsDUO1. The MpChimera was able to rescue the activation level to a comparable level as PpChimera, which was expected given the similar sequences of the two in the MYB domain. The GAChimera, on the other hand, showed no evidence of target activation.

This indicated not only the MpDUO1 MYB domain recognise and bind to the same target sequence as AtDUO1, but the role as an activator for MpDUO1. The weak activation ability of MpDUO1 was also shown *in planta* when transformed into *A. thaliana* plants. It was able to rescue the cell division phenotype fully, but showed no

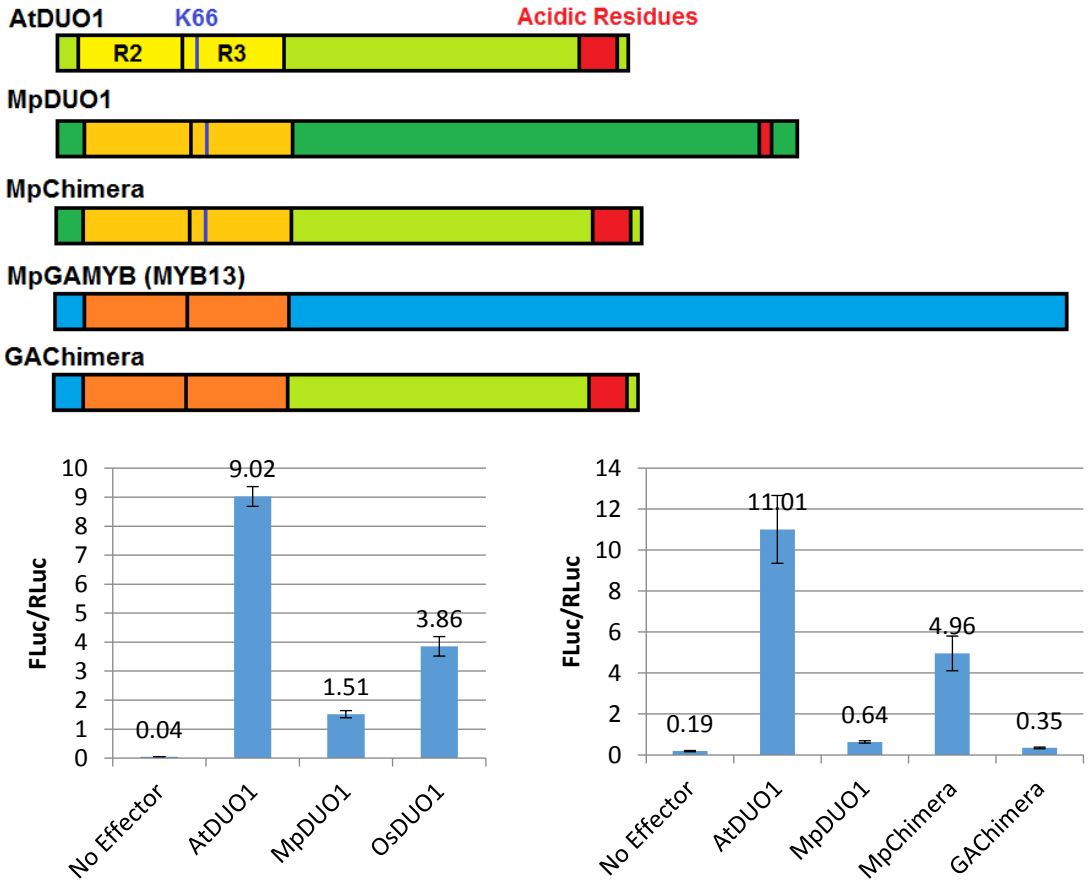


Figure 4. 8 Chimeric molecules were made from MpDUO1 and MpGAMYB (MYB13), fused with the AtDUO1 C-terminal region.
The approach is the same as for the PpChimera. Molecules were made using PCR, then put into tobacco leaves for transactivation tests. MpDUO1 shows a weak ability to activate the *A. thaliana HTR10* promoter. MpChimera was able to increase the target promoter activity into ~50 % of the AtDUO1 positive control, comparable to the PpChimera. The GAChimera did not show any activation capacity. Yellow/Orange = MYB domains.

rescue of the cell differentiation phenotype (Higo, unpublished).

As discussed in Chapter 3, the key differences between a DUO1 MYB domain and a GAMYB MYB domain were identified in three regions. Therefore, a series of hybrid mutants were created with site-directed mutagenesis (Figure 4. 9).

In Construct 2-4, the Region A, B, and C were switched to the sequences of GAChimera MYB sequences, respectively. In Construct 5, both Region B and Region C were switched, and all three regions in Construct 6. Construct 8-12 contained reciprocal switches from GAChimera to MpChimera sequences. These constructs helped to gain further understanding in which regions were critical for the DUO1 target sequences recognition and binding.

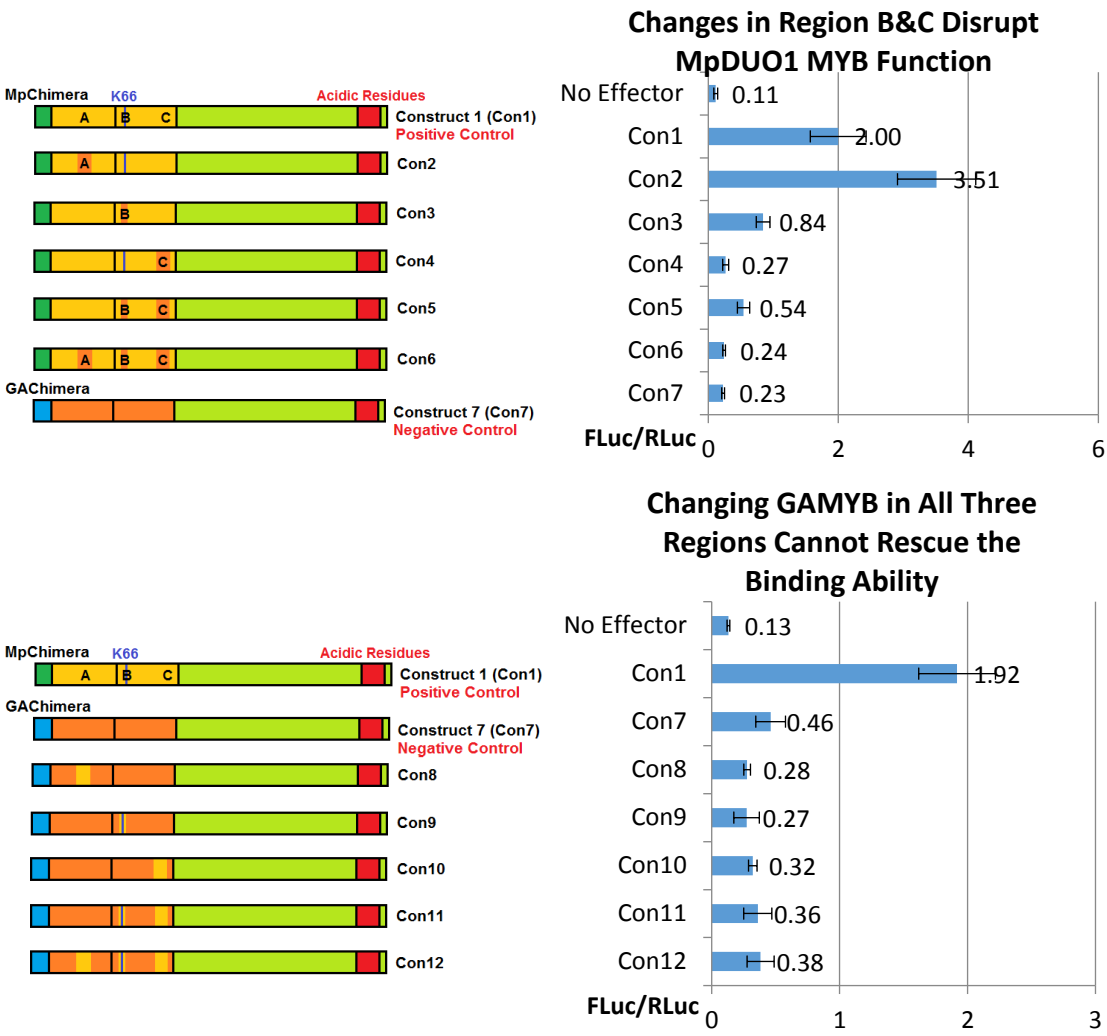


Figure 4. 9 Hybrid molecules were constructed using site-directed mutagenesis. Construct 1 and Construct 7 are the MpChimera and GAChimera, used as the positive control and negative control, respectively. Construct 2-4 and 8-10 all contain one region that has been switched to the opposite sequence (MpChimera to GAChimera or *vice versa*). Construct 5 and 11 have both Region B and C switched. Construct 6 and 12 have all three region switched. Changes in Region A and C involves only substitutions of residues, while changes in Region B create a 2-amino-acid length difference (MpChimera to GAChimera is -2, and the reverse is +2). Construct 2-4 contains single region switch of Region A, B, and C. Con4 had the most reduced activity out of the three, Con3 also showed a huge decrease. Con2 showed a significant increase in its target activation after the disruption of the Region A. Combination of the changes reduced the activation ability to a level similar to Con4. The reciprocal constructs designed to increase the activation ability of Con7 did not behave as expected. None of the constructs showed any increase in the activation level. Boxes indicate switched regions.

These constructs were then tested in the transient assays to activate the *A. thaliana HTR10* promoter. Construct 2-6 were tested to see the detrimental effects on the disruptions of the three regions. Construct 8-12 were also tested for attempts of activation ability restoration.

For the single region switch variants, disruption in Region A (Con2) not only failed to have a negative effect, but showed an enhancement of the activation ability. This was possibly because of the mutation destroyed the miR159 binding site, which is discussed in the next chapter. Switch of the Region B (Con3) had a strong negative effect on the activation level. However, since this region is not directly involved in DNA recognition, this could be the result of the protein length change between the R2 and R3 repeat. The loss of the two amino acids might have affected the flexibility of the MYB domain, which in turn affected its DNA binding ability. The change in Region C (Con4) showed the greatest impact on its target activation ability. This was not surprising since it forms the third α-helix, which is crucial for the MYB target sequence recognition. The double and triple switch had similar reductions on target activation level as Con3 (change in Region B). However, they retained a low level of binding ability, similar to the mutation to the W86G mutation, which a core residue of the R3 repeat is replaced (Borg, 2011).

However, none of the changes made was able to increase the target activation ability of the GAChimera, which was dictated by the MYB domain binding ability towards the DUO1 target sequence. This suggests that there are other sequences or residues that contribute to the target sequence difference of the two proteins, even though the three regions contain most of the differences in the MYB domain between the DUO1 and GAMYB.

Assuming the remaining length of the region between the R2 and R3 domain still provides enough flexibility after the switch of Region B in the GAMYB, the binding site recognition is mainly dependant on the third α-helices of the two domains. Since the helix of R3 (Region C) was replaced in the Con5 & 6, the most probable cause for the binding failure was the difference in the third helix of the R2 repeat (Figure 4. 10).

This is the region where the R2 repeat interacts with the target DNA as mentioned. Although only two amino acid changes exist here, i.e. from alanine to valine (AtDUO1 V54) and histidine (side chain pKa 6.0) to lysine (side chain pKa 10.5, AtDUO1 K56), and they would result in only a limited shift in the polarity of this region, the structures of the side chain might also play a role in the DNA recognition. The H → K switched an aliphatic side chain to an aromatic one, while the A → V increased the size of the side chain. These two amino acids adjacent to the core tryptophan W53 are critical and must be properly placed (Saikumar et al., 1990, Heim et al., 2003).

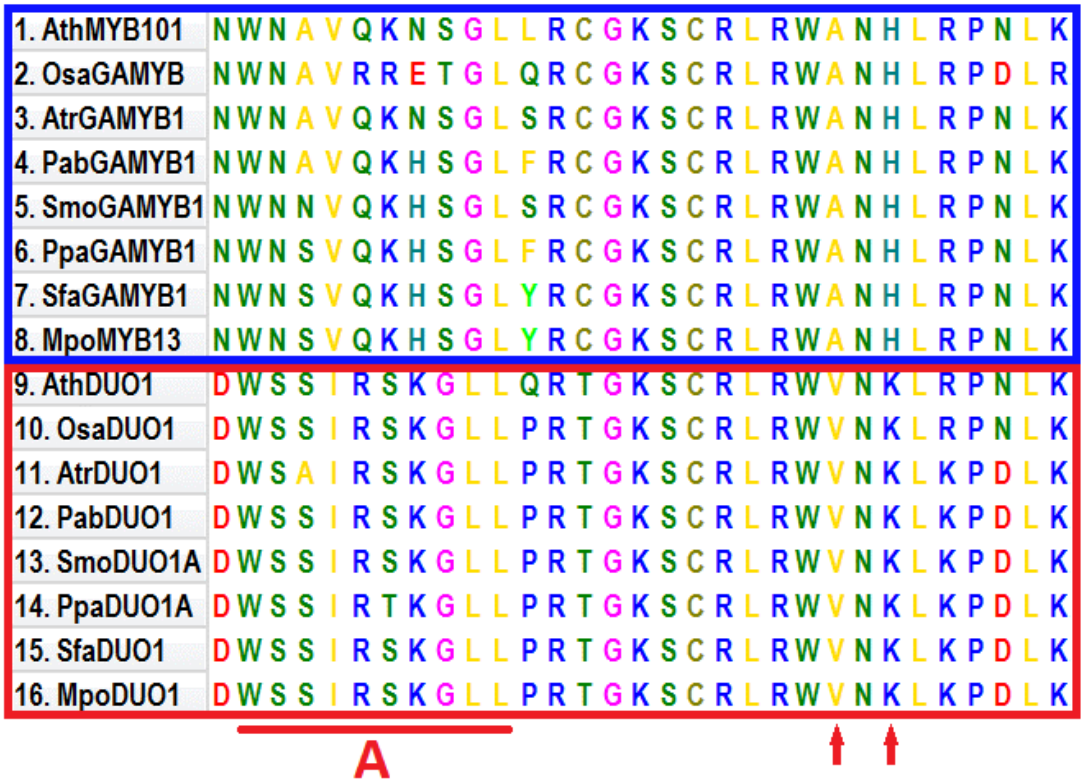


Figure 4. 10 The third helix of the R2 repeat may carry key residue changes responsible for the binding failure for the constructs.
Potential targets are labelled with the red arrows. Although these changes do not largely affect the polarity of the region, they do have huge differences in their amino acid side chains. Blue box = GAMYB; Red box = DUO1.

In a different approach, Dr JM Franco-Zorrilla from the Centro Nacional De Biotecnologia in Spain were successful in *in vitro* characterisation of the DNA binding sequence specificity for some of these constructs (Figure 4. 11). These results showed the details of how the small changes in these regions altered the target binding sequences. For the AtDUO1 and MpDUO1, the binding sequence were the same G/AACGGTTA, which is the reverse complementary sequence to TAACCGT, as predicted by Dr Borg (Borg et al., 2011). The target sequence for the GAMYB protein, MpMYB13, was not conclusive. Neither was the binding sequence for the *K. flaccidum* R2R3 MYB homolog. Although for MpMYB13, the second highest scoring motif has the same binding sequence as the highest scoring motifs for the Construct 4-6.

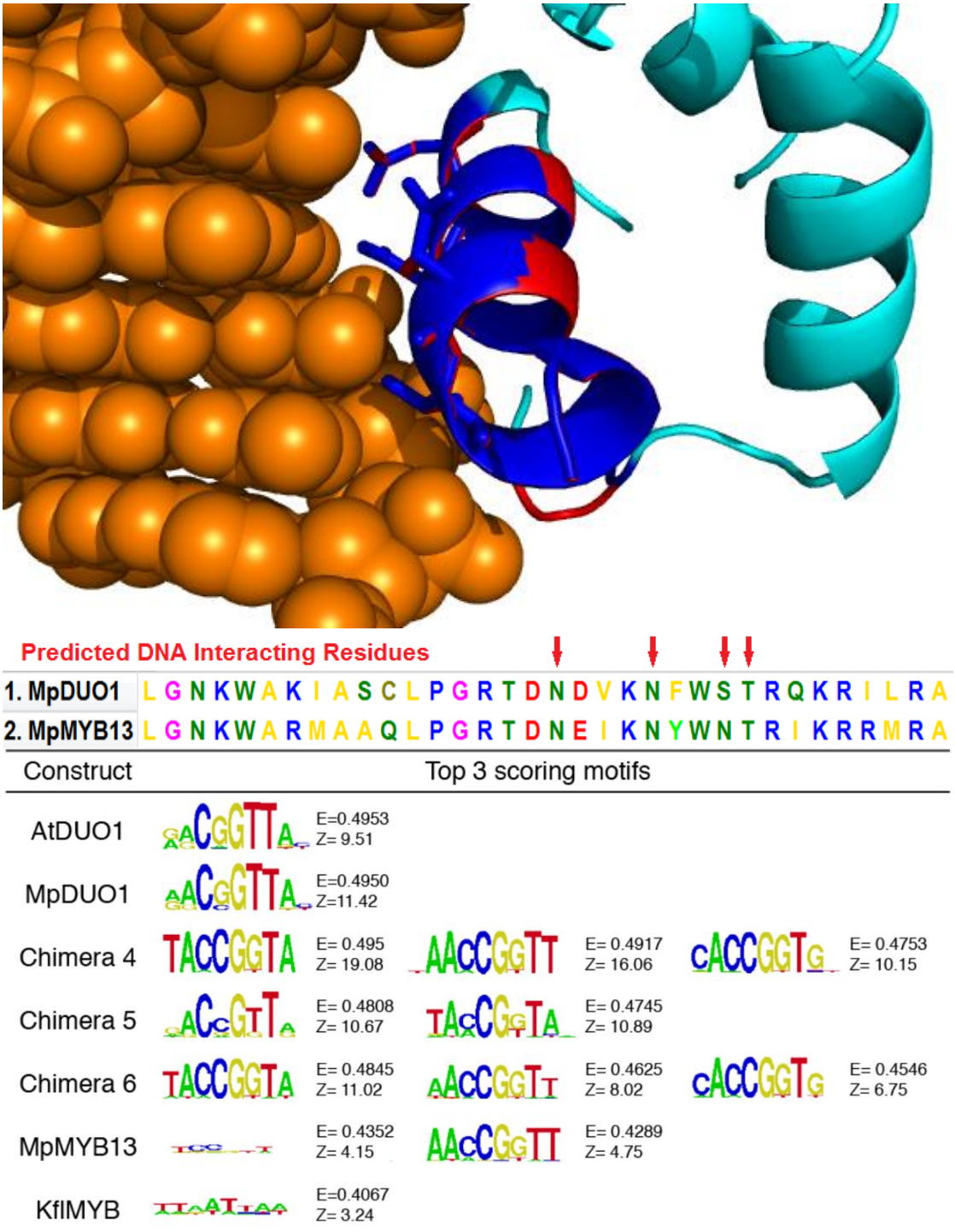


Figure 4. 11 Region C, which located in the third α-helix of the R3 repeat, is the responsible for the target sequence recognition.
This helix sits in the major groove of the target DNA (orange). Changes of residues here will very likely change the target sequence. The residues shown in sticks (blue = GAMYB; red = DUO1) and marked with red arrows in the alignment are predicted to interact with the target DNA (Solano et al., 1997). The DNA binding specificity of some MYB domains from Figure 4.9 were tested by Dr. Franco-Zorrilla, through position weight matrices of the top 8-mer DNA sequences bound on a protein binding DNA microarray (Franco-Zorrilla et al., 2014).

As discussed earlier, Region C is required for the target recognition, and the key residues responsible can be predicted (Solano et al., 1997). Although there was only one such key residue change (S → N), the binding sequence of the Construct 4-6 changed from AACGGTTA to TACCGGTA, a similar but definitely different one.

With these results combined, the current model is that Region C, which forms the third helix of the R3 repeat that directly interacts with its target DNA, contains a key change that separated the DUO1 clade and GAMYB clade a long time ago. Since then, the MYB domains of all DUO1 homologs have kept the same function and bind to the same sequences. All consequent changes in the molecules were the results of changes in the C-terminal region of the proteins. This idea is further discussed in the Chapter 6.

4.6 Discussion

DUO1 and the GAMYB proteins evolved from a common ancestor R2R3 MYB protein in the fresh water green algae. Although both clades are involved in the male reproduction, GAMYB homologs are not male germline-specific, unlike DUO1.

DUO1 homologs have been functionally linked to reproduction male germline development since at least bryophytes. They assumed much more specific roles after the emergence of the angiosperms. Evidence suggests that DUO1 homologs from this clade are able to replace one another and still fully support the transcription network required for the germline development.

The function of a DUO1 homolog can be tested with either the *in planta* complementation or the *in vivo* dual-luciferase transient assays. Both methods provide useful information as shown in the results above. However, they also have their drawbacks. For *in planta* complementation, when the transgenes failed to express, such as in the cases of DUO1 sequences from *Solanum lycopersicum* (A), *Physcomitrella patens* (A and B), and *Amborella trichopoda*, no conclusion could be drawn from the experiments. This method is also time consuming, requiring two generations of plants. For the transient assays, the activation of a target promoter does not necessarily translate into a functional network in the native context, or guarantee a stable expression for the protein itself.

Bryophyte DUO1 homologs are not functionally replaceable with angiosperm DUO1 homologs, or *vice versa*. This is possibly due to the different regulatory networks

between the two groups. Nevertheless, the MYB domains from the two groups can still recognise and bind to the same target sequence. The MYB domain contains some important changes in comparison with the GAMYB protein MYB domain.

One of the key region for the target sequence specificity is located at the third α-helix of the R3 repeat, a few residue changes in the region would result in a target sequence change.

Another change happened between the R2 and R3 repeat, and became the DUO1 signature residue, Lysine 66 (K66). If this residue and the adjacent two residues were replace to a GAMYB residue, the target sequence would also change. However, it is worth noting that in this design, there is a two-amino-acid length change.

Previously, the K66 was deleted or mutated for AtDUO1, but no such dramatic result was achieved (Borg, 2011). In fact, the deletion of the K66 (ΔK66), which shortens the protein length by one amino acid, increased the activation ability of DUO1 *in vivo*. Considering MpDUO1 is one amino acid longer than AtDUO1 (an extra arginine), the ΔK66 has the same length as the Construct 3, thus should behave similarly. It remains unclear why one has a stronger binding affinity towards the original target and the other one changed its target sequence. There are sequences outside the three examined regions that are contributing to the target sequence specificity. It is possible the changes in these sequences between the MpDUO1 and AtDUO1 were responsible for the different behaviours of ΔK66 and Con3.

The last change examined was not required for the sequence recognition. Instead, it became a microRNA-binding site in the core angiosperms. However, the role of the miR159 on DUO1 remains a speculation.

Chapter 5 The microRNA Regulation of the DUO1 and GAMYB Clades

Abstract

Although it has been reported that AtDUO1 contains an miR159 binding site, and the miR159 regulation of DUO1 expression has been independently observed many times during ectopic expression experiments, no miR159 regulation inside the germ cells has been reported so far. The miR159 binding site is located in a region where significant amino acid changes have occurred during the emergence of DUO1, which separated this clade from the GAMYB proteins. However, miR159 does not exist in the bryophytes, only its relative miR319. Binding site positions and mismatch comparison evidence indicated that the miR159 regulations were recruited separately in the *DUO1* and *GAMYB* clades. By comparing the DUO1 binding sites in different clades from bryophytes to angiosperms, and the miR159/319 family history, a hypothesis was proposed that the gaining of the miR159 binding site was a random event through genetic drift, and subsequently losing this site could be a nearly neutral mutation that occasionally gets fixed in some species.

5.1 Introduction

Binding sites for miR159 have been identified in the transcripts of both *DUO1* and *GAMYB* in *Arabidopsis thaliana* (Jiang et al., 2004). Sequence comparisons in Chapter 3 revealed that the miR159 binding sites for the two clades were located in different parts of the genes. This strongly suggested that the regulation of these two groups by miR159 might have developed independently.

Detailed evidence have been collected for the regulation of the *GAMYB* members through miR159, as discussed in Chapter 1. The precursor of miR159 in the bryophyte *Marchantia polymorpha* miR319 has also been reported to cleave the target *GAMYB* mRNA (Tsuzuki et al., 2016). Information about either the miR159/319 or the *DUO1*/*GAMYB* in Charophyta algae remains elusive. No credible miR159/319 homolog has been found in Chlorophyta (Zhao et al., 2007, Li et al., 2014c).

On the other hand, most of the evidence for the miR159 regulation of *DUO1* was based on ectopic expression. The miR159 resistant form *mDUO1* caused a severe phenotype when expressed using the *LAT52* promoter, but not the native form *DUO1* (Brownfield et al., 2009a). An independent study using the 35S promoter also came up with similar results, with mild defects in 33 % of the T1 plants caused by *DUO1*, but severe defects and sterility in 40 % of T1 and mild defects in most of the remaining T1 plants by *mDUO1* (Palatnik et al., 2007). Based on these observations, it was proposed that miR159 regulation was important for the normal function of *DUO1*.

However, *DUO1* is normally expressed exclusively inside the male germline cells and has a conserved regulatory region in its promoter (Peters et al., 2016). The pollen-specific ARID1 was shown to bind to the promoter region of *DUO1* in the ChIP assay and positively regulate its expression in BCP and TCP (Zheng et al., 2014). It was proposed that miR159 and ARID1 co-regulate the expression of *DUO1* in the vegetative and generative cells but so far this model remains speculative, considering no expression feedback loop of the two was reported in the study.

Curiously, the DUO1-RFP signal and the *DUO1* transcript was only reduced to around half of the normal level in *arid1-1*. An expression difference was also observed for the DUO1 protein variants using miR159 sensitive/resistant forms by Dr Michael Borg. It was later quantitatively measured and presented in this chapter.

5.2 The miR159 clade evolved from the miR319

Since both *DUO1* and *GAMYB* homologs in *A. thaliana* are regulated by the microRNA miR159, which is closely related to the ancestral miR319 family (Palatnik et al., 2007, Li et al., 2011b), the miR159/319 evolution in plants becomes very important.

There are six members of the miR159/319 clades in *A. thaliana* (3/3), five in *P. patens* (0/5, no miR159 exist), and only two in *M. polymorpha* (0/2) (Tsuzuki et al., 2016). The two MpmiR319 homologs have an identical sequence and were proven to target the *AtMYB33* homolog in *M. polymorpha* (Tsuzuki et al., 2016, Lin et al., 2016). Three of the five miR319 homologs from *P. patens* also have the same sequence (Axtell et al., 2007, Axtell and Bowman, 2008). All these sequences were aligned (Figure 5. 1). In *A. thaliana*, the three miR319 homologs had a 1 bp shift compared to the miR319 homologs in bryophytes, but the core sequences remained the same. The miR159 homologs had no shift compared to the bryophyte miR319s, but contained changes of nucleotides at both ends and one in the middle.

Information about miR159 and miR319 homolog numbers from different plant clades were collected from the website miRBase (http://www.mirbase.org/) (Griffiths-Jones et al., 2006, Meyers et al., 2008, Kozomara and Griffiths-Jones, 2014). The miR159 does not exist in bryophytes, but has a copy in *S. moellendorffii*. This suggests that miR159 and miR319 most likely separated from each other in lycophytes.

Although the copy numbers of the miR159/319 homologs in different species fluctuate, it is important to note that in the case of *A. thaliana* miR159, the multiple copies have functional redundancy (Allen et al., 2007).

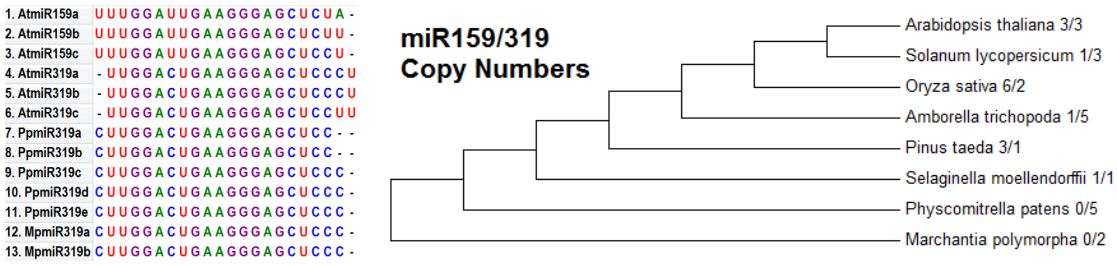


Figure 5. 1 The two microRNA clades miR159 and miR319 are closely related.
The miR319 homologs in bryophytes have the same sequence in both *M. polymorpha* and *P. patens*. They evolved into miR159 and miR319 two clades in lycophytes like *S. moellendorffii*. In *A. thaliana*, the miR319 sequences remain the same but have a 1bp shift. The miR159 sequences has no position shift, but contain several nucleotide changes, mainly C → U.

5.3 The miR159 regulation depends on sequence complementarity

Like the other microRNAs, the regulation abilities of miR159 on its targets can be directly affected by mismatches (Li et al., 2014b). Artificial versions of its targets altered at the binding sites have been shown to avoid its cleavage effect, like *mMYB33* (Allen et al., 2007, Li et al., 2014b), *mMYB101* (Allen et al., 2010), and *mDUO1* (Palatnik et al., 2007, Brownfield et al., 2009a).

It is not yet clear if miR159 regulates *DUO1* expression in the germline cells. However, there is evidence for the regulation ability of miR159 on *DUO1* in leaf cells. A previously mentioned example of the *mDUO1* in *A. thaliana* has been shown to cause severe phenotypes when ectopically expressed (Brownfield et al., 2009a). In transient assays, mDUO1 also shows a significant increase of its target activation levels compared to the miR159 sensitive native form of DUO1 (Borg, 2011).

Many angiosperm *DUO1* homologs contain an identical sequence at the potential miR159 binding site, like *Oryza sativa*. However, some *DUO1* sequences analysed contained enough mismatching to the miR159 that they were predicted not to be regulated, like the monocot *Lilium longiflorum*.

To confirm that the *LlDUO1* is not under the regulation of miR159, a sensitive mutant form *sLlDUO1* was created with the help of Mikhaela Neequaye (Figure 5. 2). In the exact opposite manner of making *mDUO1*, three nucleotides were mutated to ensure the matching between this artificial molecule and the miR159. Both versions of the LlDUO1 and AtDUO1 were tested in the transient assays.

The mAtDUO1 had an increase in activation of target promoters compared to the normal AtDUO1, confirming previous reports. The native form of LlDUO1, which does not contain a miR159 binding site, had a much higher activation level of the target promoters than the AtDUO1. The increase of its susceptibility to the miR159 caused a reduction in its activation abilities. The tests were performed on three different DUO1 target gene promoters, including *HTR10*, *DAZ1*, and *DAA1*. They all showed a similar trend for these constructs.

This indicates that *DUO1* homologs are not always susceptible to the miR159 regulation, even in vegetative tissues. Within the core angiosperm group, most species have the miR159 binding site on their *DUO1* sequences. Have they acquired this

feature separately in multiple branches, or did it evolve in a much earlier ancestor but some species have lost it later? The first step to answer this question would be to find out the history of the relationships between the miR159/319 and *DUO1*.

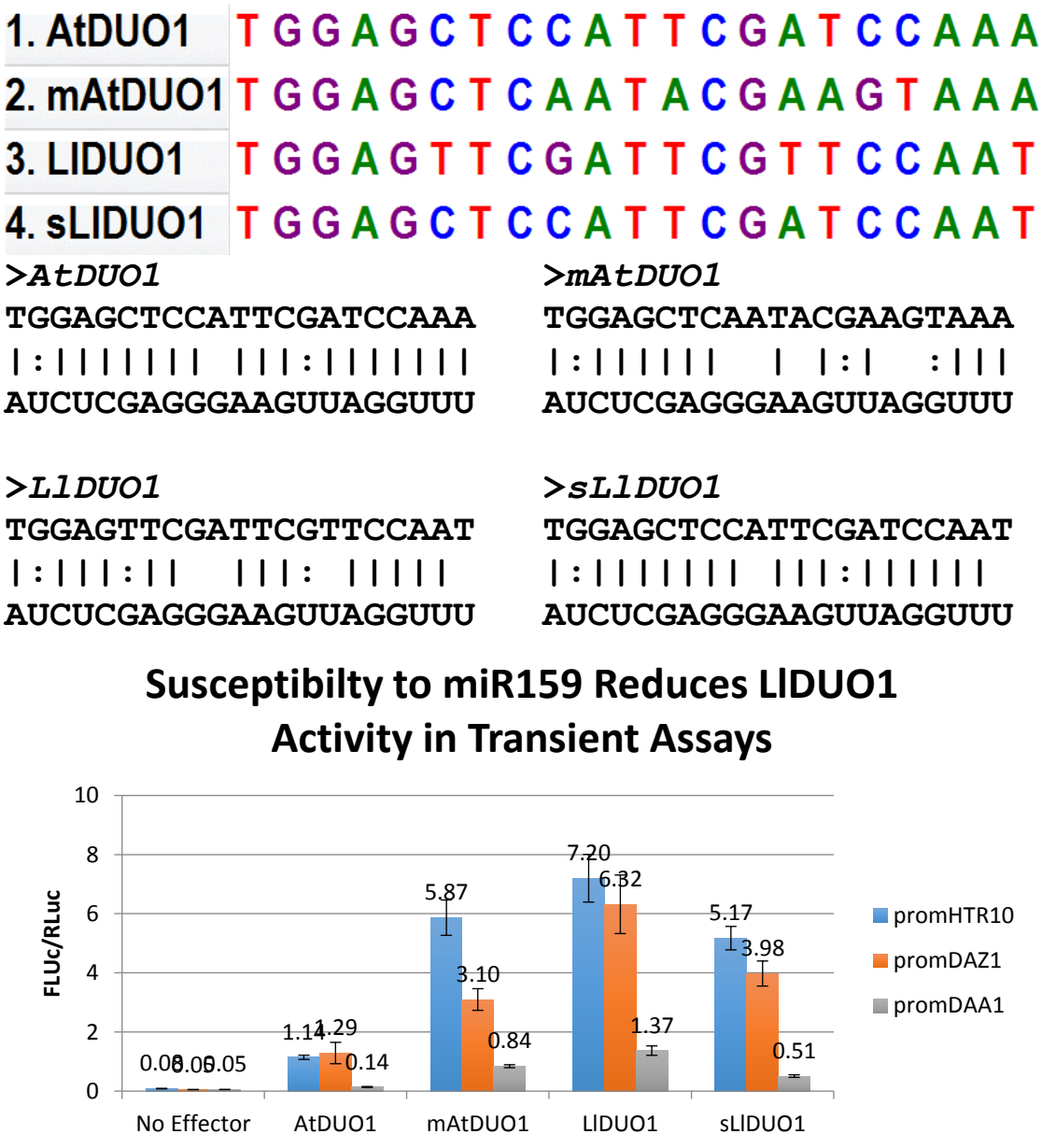


Figure 5. 2 Not all DUO1 homologs are susceptible to the miR159 regulation. Increasing the mismatches between the miR159 and *DUO1* in *A. thaliana* will make it resistant to the microRNA regulation. On the other hand, increasing the matches between the miR159 and *DUO1* in *L. longiflorum* can turn it from resistant to susceptible to the miR159. There are a number of *DUO1* homologs outside and within the angiosperm group that do not contain a sequence matching for the miR159. They are predicted not to be regulated by it. Transient assays were performed by Mikhaela Neequaye.

5.4 *DUO1* and *GAMYB* clades recruited miR159 separately

Like other microRNAs, the miR159/319 clade evolves at a very slow rate. The only copy of miR159 in *Amborella trichopoda* has an identical sequence to the miR159a in *Arabidopsis thaliana* (Albert et al., 2013). These microRNA sequences start to become different outside the angiosperms, into the gymnosperms (Lu et al., 2007), or the lycophytes (Axtell et al., 2007). However, this rate of change cannot account for the relatively rapid changes in either the *GAMYB* or the *DUO1* homologs.

Although both can be regulated by the miR159 in *A. thaliana*, the binding site positions of the *GAMYB* genes and *DUO1* are different. For the *GAMYB* clade, the binding site is outside of the MYB domain, but for *DUO1* it is within the MYB domain which was labelled as Region A in Chapter 3. Therefore, until new evidence suggests otherwise, it is reasonable to assume the miR159 binding sites for the *GAMYB* clade and *DUO1* clade are not evolutionarily homologous.

The *GAMYB* family was under the regulation of miR319 in the bryophytes (Figure 5. 3). In *Marchantia polymorpha*, the miR319 was able to cleave its target *GAMYB* DNA *MpMYB13* with four mismatches (Tsuzuki et al., 2016). As discussed previously, the miR159 separated from the miR319 in the lycophytes, and are known to be responsible for the regulation of the *GAMYB* genes in some angiosperms. All the *GAMYB* sequences that have been collected so far contain the miR159 binding site that has the same number of the mismatches as the *MpMYB13* or less.

This does not automatically imply that all *GAMYB* genes are directly regulated by miR159. For instance, the resistant form *mMYB101* did not cause any phenotype in anther or pollen (Allen et al., 2010). Previously it has been reported that miR159 does not regulate its targets (e.g. *MYB33/65*) in vegetative tissues as it does in certain special tissues (e.g. seeds) (Alonso-Peral et al., 2012).

On the other hand, *DUO1* homologs were not likely to be regulated by miR159 until angiosperms (Figure 5. 4). Even in the basal angiosperm *A. trichopoda*, the supposed miR159 binding site contains too many mismatches to the sequence. This was the case for all the earlier divergent species examined so far. Most species within the core angiosperms contain the identical sequence for the miR159 binding. A few *DUO1* sequences, like *LlDUO1*, were predicted to be resistant.

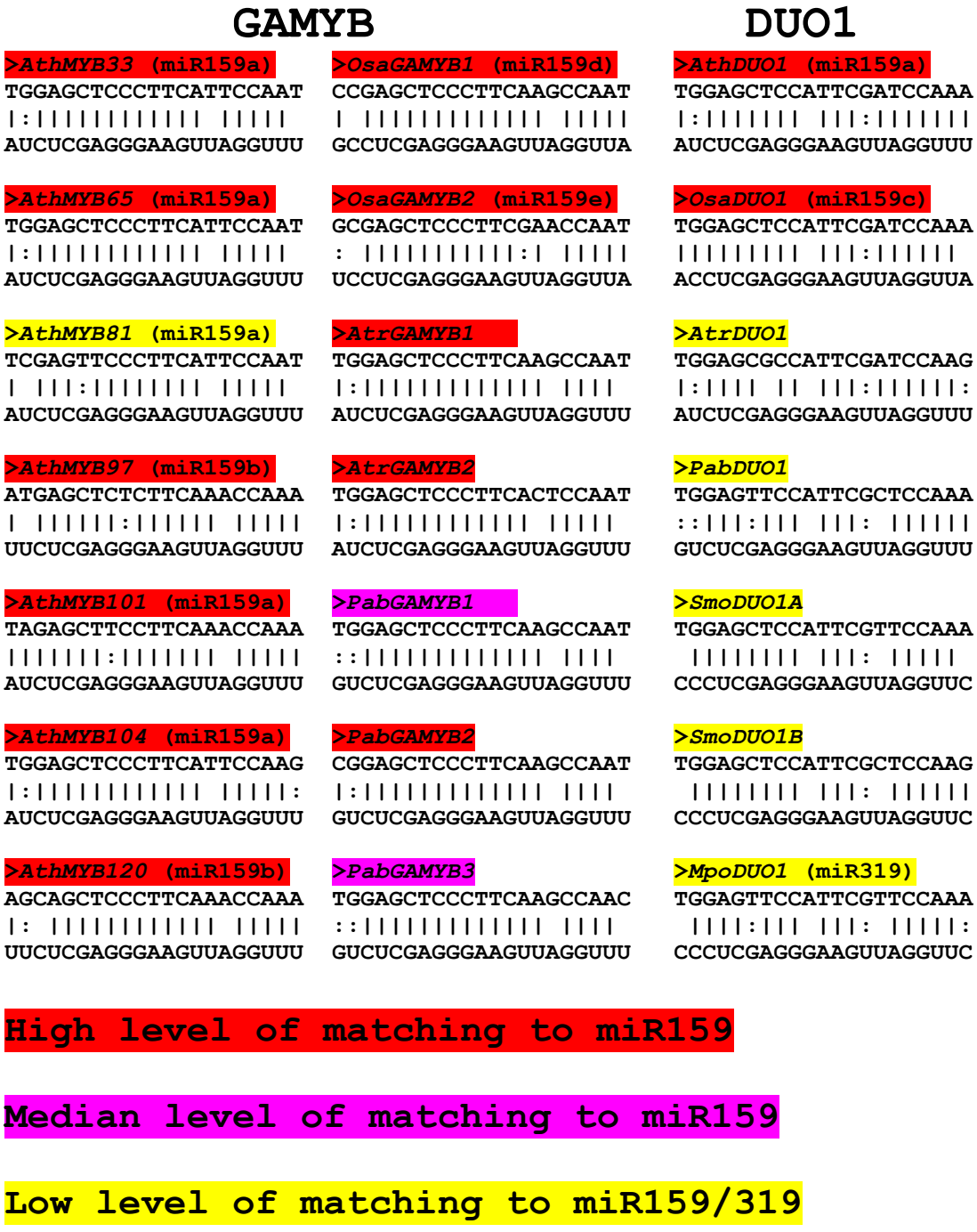


Figure 5. 3 *GAMYB* homologs have been under miR159 regulation before angiosperms, while *DUO1* homologs were only susceptible to miR159 regulation after the establishment of the core angiosperm group.
In *M. polymorpha*, miR319 is able to cleave the target *GAMYB* with four mismatches. This level of matching between the *GAMYB* homologs and miR159/319 continued until the gymnosperms. The matching level has been increased in the angiosperms. For *DUO1* homologs, the miR159 binding sites, which locate inside the MYB domain, do not have the same origin as the ones from the *GAMYB* homologs. Even for the basal angiosperm *A. trichopoda*, the mismatch level between the *DUO1* and miR159 are too high for any effective regulation. Parentheses indicate the microRNA with the maximum matching when multiple copies exist in the species; miR159/319 sequences were collected from miRBase.

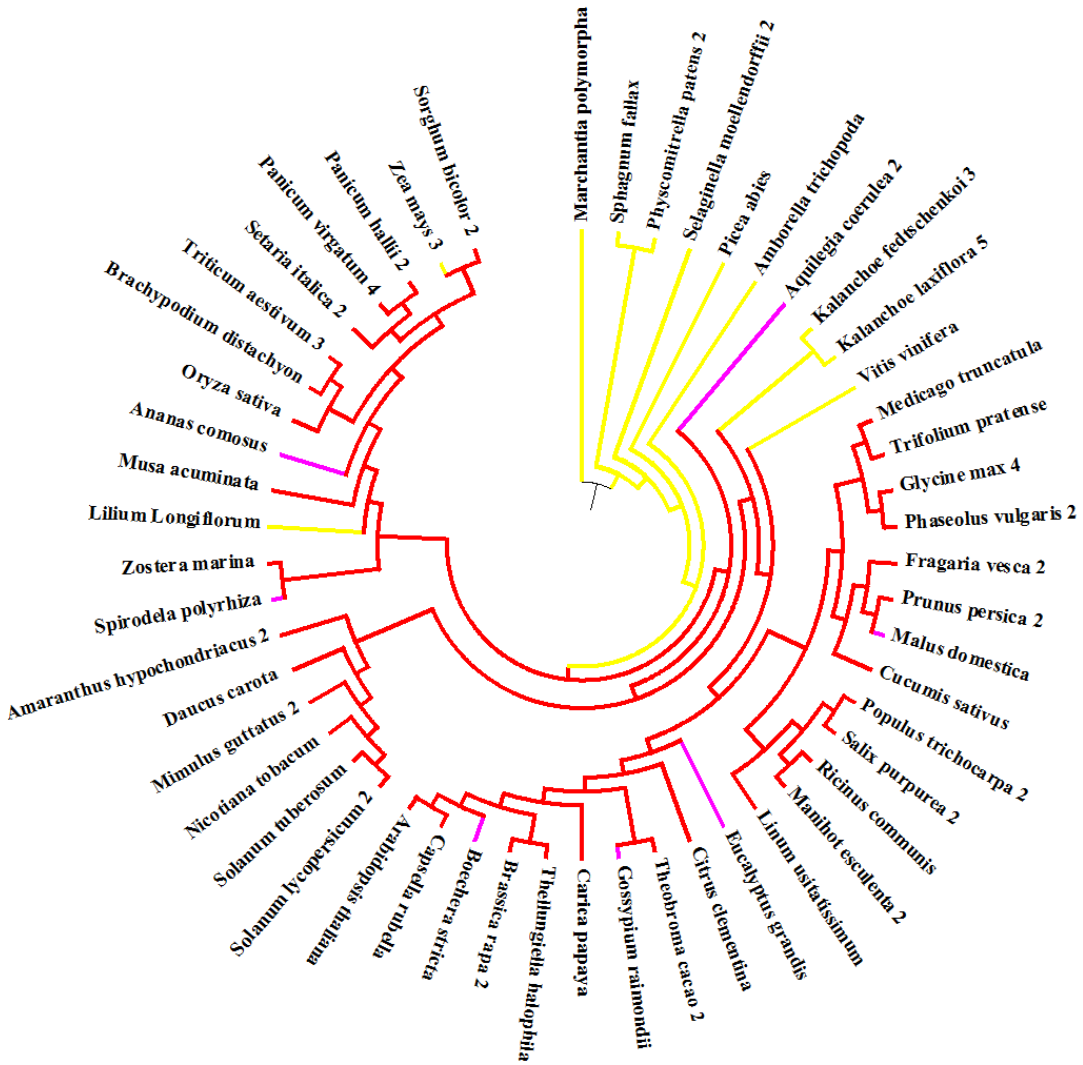


Figure 5. 4 The interaction between the *DUO1* homologs and the miR159 was developed in the core angiosperm group, but the regulation is not always required.
Although absent in the basal angiosperms and all groups before that, many species from both the eudicots and monocots have an identical miR159 binding site, suggesting this was evolved in the common ancestor of the two groups. Some species from different sub-clades contain mutations that disrupt the binding site. The cladogram was built using the same sequences of *DUO1* described in Chapter 3.

In Chapter 3, it was discussed that a key structural changes had already happened in the MpDUO1 MYB domain (Region A). Although this region from the two species looked similar structurally, one is susceptible to miR159 in *A. thaliana* but not the other to

miR319 in *M. polymorpha*. In Chapter 4, it was also demonstrated that this region was not critical for the MYB DNA target binding and recognition functions.

Combining these results, it seems that the changes happened in the MpDUO1 Region A was a mutation unrelated to the microRNA regulation. *DUO1* recruited the miR159 for its regulation after the establishment of the core angiosperm group. The abundance of the resistance mutation in separate clades indicates this being possibly a type of slightly deleterious mutation (Ohta, 2011). In other words, the sequence variation in this site predicts that the advantage provided by the miR159 regulation on *DUO1* is minute. Species with a small population size (e.g. speciation events) have a chance to lose the binding site without being purged by selection, and consequently fix the mutation through genetic drift. However, the mechanism of the fitness change is unclear.

5.5 The *in planta* regulation of miR159

The expression of *DUO1* is restricted and only inside the germline cell nucleus (Chapter 1 and 3). If the miR159 regulation on the *DUO1* expression exists in this context, however subtle, it should be possible to measure the differences.

Two constructs were made using the native form of *DUO1* and the resistant form of *mDUO1* from *A. thaliana*. They were attached to a red florescent protein "mCherry" and driven by the *DUO1* promoter, namely pB7m34GW_promDUO1:AtDUO1-mCherry and pB7m34GW_promDUO1:mAtDUO1-mCherry. They were then transferred into regular wild type *A. thaliana* ecotype Col-0.

The first direct observation was done by Dr Michael Borg who reported that the resistant form mDUO1 in plants appeared to have a lower expression level, contrary to the prediction that miR159 would cleave the sensitive native form of *DUO1* and make it express at a much lower level. To confirm the observation, the experiment was then independently repeated, starting from making the two constructs. The expression levels were then measure indirectly through quantifying the mCherry florescence signal levels under the microscope (see Materials & Methods chapter).

The experiment was repeated twice, and they yielded similar results. After transforming the constructs into the plants, a puzzling phenomenon was observed (Figure 5. 5). The plants that carried the *mAtDUO1* transgene had an unusually high mortality rate after

being transferred into individual pots (60 out of 80 died), while the ones carrying the *AtDUO1* transgene were unaffected (2 out of 80 died). Most of the death occurred during the principal growth stage I, when vegetative rosette development was gaining leaf numbers (Boyes et al., 2001). In other stages, these individuals could not be morphologically distinguished from the wild type plants. Although there were fungal infections for the *mAtDUO1* lines, the individuals selected for transferring did not appear to be infected, and the death did not occur at an earlier stage when the seedlings were more vulnerable to an infection.

Same numbers of plants for each transgene were then collected and pooled together for the quantification of the protein expression. The results agreed with the initial observation from Dr Borg. In both experiments, the miR159 sensitive AtDUO1 was expressed at about twice the level of the miR159 resistant mAtDUO1 (Figure 5. 6).

Although unlikely to be the reason, the three synonymous nucleotide changes might have an impact on the expression. An experiment has been designed to test that idea. The same two constructs were made with a new antibiotic (Hygromycin) selection marker: pH7m34GW_promDUO1:AtDUO1-mCherry and pH7m34GW_promDUO1:mAtDUO1-mCherry. They were transformed into the triple homozygous mutant *miR159abc A. thaliana* lines (Allen et al., 2007, Palatnik et al., 2007). If these two proteins are expressed at the same level, it would show that the interaction with the microRNA was the reason for the difference observed in these experiments.

However, given the expression level difference was preceded by the mortality rate difference, which had never been associated with codon optimisation previously, in this thesis it is assumed that the microRNA regulation was responsible for the observation.

These results suggest that the miR159 regulates *DUO1* in a completely different manner inside the germ cells, if at all. However, even if miR159 was responsible for increasing the expression of DUO1 through unknown mechanism, a two fold level change was unlikely to be "nearly neutral" as predicted earlier. On the other hand, if the regulation of the miR159 on *DUO1* does not happen in the germ cells, then the lost of the miR159 binding site would have little effect on DUO1 *per se, ergo* "nearly neutral". That would make the resistance to miR159 not beneficial for raising the DUO1 expression level, but does not explain why the mAtDUO1 level was lower.

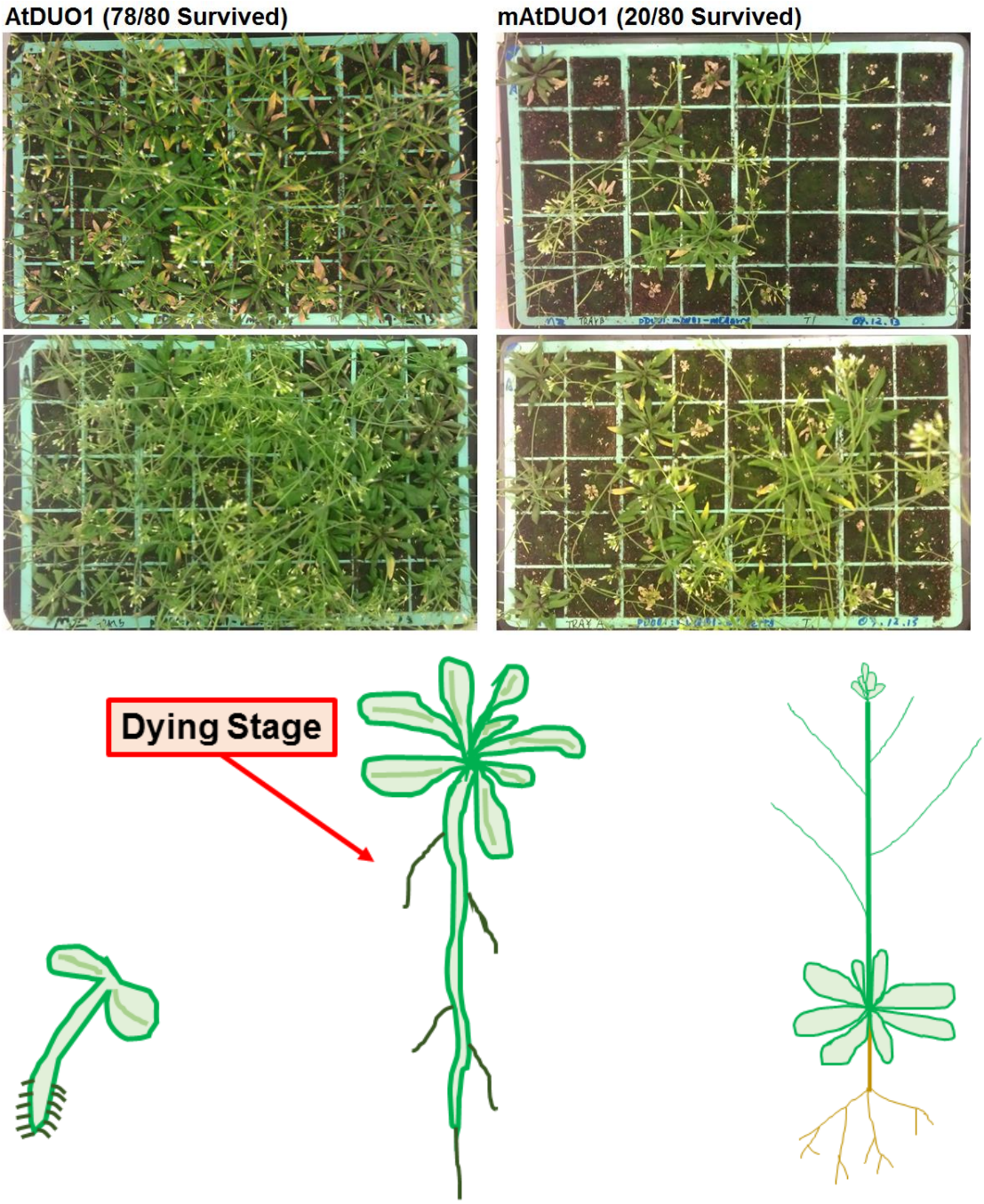


Figure 5. 5 The plants carrying the mAtDUO1 transgene had a high death rate at the principal growth stage I.
Compared to the plants carrying the wild type *AtDUO1* transgene of which most were growing normally, the mortality for *mAtDUO1* was unusually high at this specific stage. Other stages of growth or tissues were not affected. Developmental map loosely adapted from the Arabidopsis eFP browser.

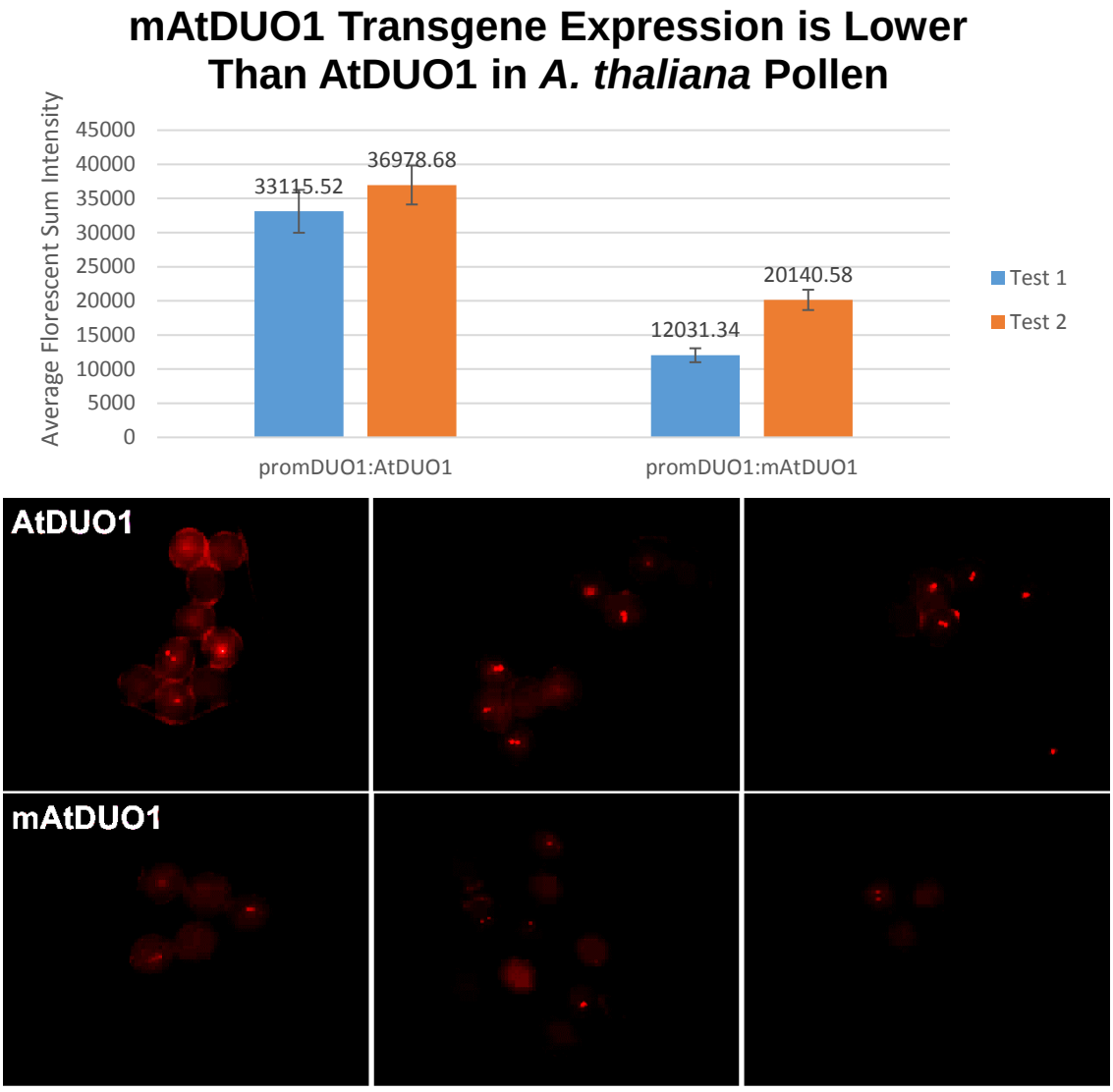


Figure 5. 6 The expression level of microRNA resistant mAtDUO1 was lower than the sensitive AtDUO1, contrary to the prediction.
The average florescent sum intensity, which represent the transcription level for the mAtDUO1 was only about half of the AtDUO1. The test was independently repeated twice and results were similar. If miR159 was reducing DUO1 transcript inside the germline cells, the results should be the opposite. Images were captured using a standard exposure time t=500 ms; error bars represent the SE; n=50; $p<0.001$.

However, correlation does not necessarily imply causation. The chance of an unknown microRNA mechanism increased a target protein expression was much lower compared to a hidden condition set caused by the miR159. In this case, seemingly separate events might have been a Markov chain started at the earliest stage through unconscious sample selections, which in turn created the statistical and logical error "survivorship bias". This is detailed in the discussion of this chapter.

5.6 Discussion

The last change on the DUO1 MYB domain examined in Chapter 4 (Region A) was not required for sequence recognition. Instead it became a microRNA binding site in the core angiosperms. However, the role of the miR159 on *DUO1* remains open to speculation.

The interaction between *DUO1* and miR159 happened after the establishment of the core angiosperm group. Incidentally, the regulatory region of DUO1 (*ROD1*) that is conserved in core eudicots was also believed to have developed around that period (ca. 140 mya) (Peters et al., 2016). *DUO1* likely gained the binding site sequence and recruited miR159 into its regulation network. This binding site however was mutated more frequently than it should have been if a high enough selection pressure was present, but it is also too conserved even for the MYB domain where synonymous mutations at the nucleotide level is common. This indicates that the gaining of the miR159 regulation is likely to be nearly neutral.

According to the nearly neutral theory, a great number of the mutations are slightly deleterious, instead of completely neutral (Ohta, 1992). Compared to strict neutrality, the slightly deleterious mutation predicts the prevalence of the low-frequency polymorphisms (Ohta, 2011). Slightly deleterious mutation can be fixed due to genetic drift in small populations, such as in a speciation event (Ohta, 1973). This is consistent with the finding as miR159 resistance was mainly found in individual species or outgroup clades, but not in big clades where population sizes were big enough to eliminate such a mutation.

This nearly neutral effect was unlikely to be the protein expression level difference observed in the miR159 resistant/sensitive *DUO1* experiment. A two fold expression level change for a transcription factor regulating both cell cycle pathway and cell differentiation pathway networks would certainly be exposed to a high selection pressure, positive or negative. Therefore the regulation is unlikely to happen inside the germ cells, but in the vegetative tissues.

This could potentially be the reason why *mAtDUO1* lines had a high mortality rate and lower DUO1 expression level. Ectopic expression of DUO1 has s catastrophic effect on the plants (Palatnik et al., 2007, Brownfield et al., 2009a). The *DUO1* promoter may not be strictly germline expressing, e.g. activated during the stress response at an early

developmental stage in the seedlings after transferred into pots. This would not be a problem normally as miR159 would silence the expression post-transcriptionally, but not if the *DUO1* is miR159 resistant. Such a "leak" in certain tissues (e.g. meristems) might have manifested in the form of seedlings death. The ones developed into mature plants were able to survive that stage due to a lower *DUO1* promoter activity, and this bottleneck effect would consequently lead to a lower DUO1 expression level in the quantification.

Dr Borg did not report that he had observed a similar phenomenon regarding the fatality of the seedlings, but healthy looking young seedlings were more favourable to researchers during the sample selection, whether consciously or otherwise. Undisturbed *mAtDUO1* line seedlings did not show a similar fate during the experiments. Therefore, it is likely Dr Borg did not trigger such a condition but inadvertently selected a similar sample population that was low on DUO1 expression due to the healthy appearance.

It is reasonable to suspect miR159 to be responsible for regulating gene expressions during stress. Its sister clade miR319 in *Arabidopsis thaliana* are linked to a wide range of abiotic stress responses (Barciszewska-Pacak et al., 2015). Another hypothesis points out that many interactions between the microRNAs and their predicted targets are not physiologically relevant. Quite the reverse, some microRNA "targets" were argued to be regulating the microRNA level, and the proposed sign of a true regulated target was its sensitivity towards the microRNA level (Seitz, 2009). It is unclear whether *DUO1* is sensitive to the miR159 level, but the *GAMYB* family could be effectively regulated if the miR159c was expressed at only 4% of the total wild type miR159 level (Allen et al., 2010). The mismatching level of the two clades are similar in *A. thaliana*, which means it is possible that *DUO1* is not sensitive to miR159 expression level and therefore not a regular target regulated by miR159.

Thus, the slight advantage miR159 binding site bring can potentially come from two aspects. The prevention of the DUO1 ectopic expression during stress responses, and the regulation of the miR159 level in the non-germline tissues. The stress-induced promoter leakage is testable using the *miR159abc* mutant lines, by extracting the RNA after inducing the possible abiotic stress, e.g. wound, drought, salinity, etc. and then perform qRT-PCR to detect any DUO1 transcript.

Chapter 6 Evolution of a Transactivation Domain in DUO1

Abstract

The C-terminal region of DUO1 was shown to be responsible for the differences in transcription activation by DUO1 homologs of angiosperms and bryophytes. It includes an acidic rich C-terminus and a highly diverse linking region that connects to the MYB domain. Sequence analysis revealed a common motif for the two groups and a clade-specific motif for each that separated the pre-angiosperms and angiosperms. Functional tests confirmed the importance of the common motif for the DUO1 transcriptional activation. The angiosperm-specific motif is required for high level activation in its native context, but this motif alone cannot increase the activation ability for the bryophyte DUO1 protein. Although the functional importance of the diverse linking region was recognised by the tests, no specific sequence or residue was identified. The characteristic sequence and ability to function in both the plant and animal kingdoms demonstrated in this chapter strongly suggest that DUO1 C-terminus is a member of the nine-amino-acid transactivation domain (9aaTAD) family. The transactivation mechanism of the AtDUO1 TAD was proposed to involve certain mediator complex.

6.1 Introduction

In Chapter 4 it was discussed that the C-terminal region of AtDUO1 contains sequences that are required for the activation of its targets. The entire *Marchantia polymorpha* DUO1 C-terminal region was found to be sufficient to support germ cell division in *Arabidopsis thaliana*, but was not sufficient for complete sperm cell differentiation, including the activation of DUO1 target promoters, such as that of the male germline-specific histone H3 variant (H3.10) termed *HTR10*. Further, the exchange of the *A. thaliana* DUO1 C-terminus for that of its counterpart from *M. polymorpha* was sufficient to support DUO1 function in the germline of *A. thaliana*. The sequence conservation and functional requirements of AtDUO1 and MpDUO1 C-termini were investigated.

These data were consistent with the requirement for a region at the AtDUO1 C-terminus for the transcriptional activation of its targets, which was previously described in an analysis of a series of deletion variants of AtDUO1 (Borg, 2011). The largest deletions removed 185 and 175 amino acids from the C-terminal end and neither retained any transcription factor activity. The shortest deletion, DUO1ΔC3, removed the terminal 25 amino acids, which possess 36 % acidic and 48 % hydrophobic amino acids respectively. This coincides with features of some known activation domains (Triezenberg, 1995). The DUO1ΔC3 variant fully complemented the failure of germ cell division of the *duo1-1* mutant, but the rescued sperm cells failed to activate the *HTR10* promoter. This demonstrated the significance of the C-terminal 25 amino acids in DUO1 for the transcriptional activation and sperm cell differentiation.

There are some similar protein domains known to act as a strong activation domains in eukaryotic cells. VP16, for instance, is a very strong activator in plants (Zuo et al., 2000, Storgaard et al., 2002, Silveira et al., 2007). The mechanism of VP16 is much better understood.

6.2 Analysis of the DUO1 C-terminal sequence

To identify candidate sequences which may be involved in the transcirptional activation by DUO1, amino acid sequence conservation of at the C-terminus was examined by alignment of DUO1 homologs retrieved from 50 angiosperms.

Using WebLogo (Crooks et al., 2004), a consensus sequence showing conservation of two motifs (DxFFDDFP & DMFD) was generated (Figure 6. 1). However, this consensus was heavily influenced by the overrepresented groups like the Brassicaceae.

The pre-angiosperm DUO1 C-terminal sequences contain some differences. The pre-angiosperm DUO1 C-terminus conservation is longer than the angiosperms. When pre-angiosperm DUO1 sequences were aligned with the representatives of selected angiosperm clades, Motif II (DM/VFD) was present in all sequences, but Motif I (FxDDFP) was absent from the pre-angiosperms. However, there seemed to be another motif upstream in the pre-angiosperm DUO1 sequences.

To have the most representative consensus in both the angiosperms and pre-angiosperms, representatives from different clades were selected and separated into two different groups (Figure 6. 2). The consensus is shown using Jalview (Waterhouse et al., 2009). The angiosperm DUO1 group contains sequences from the basal angiosperm, basal eudicot, commelinid monocots, rosid, and asterid. The consensus had the same two motifs. While the pre-angiosperm DUO1 group also had two conserved motifs, only Motif II (DVFD) seemed to be homologous to the angiosperm Motif II. The pre-angiosperm Motif I^{+} (DSCSPDSV) was not present in the angiosperm DUO1.

Noticeably, both the length (Chapter 3) and the C-terminal region of the pre-angiosperm DUO1 homologs have maintained a high level of conservation from the bryophytes (ca 500 mya) to gymnosperms (ca 150 mya).

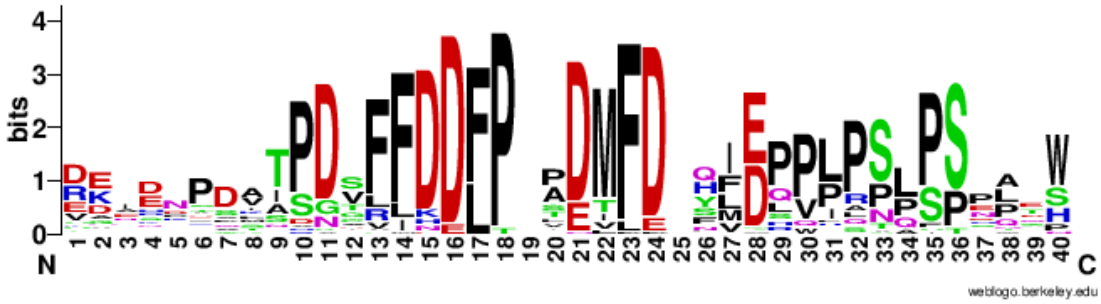


Figure 6. 1 C-terminal conservation of the DUO1 proteins.
Fifty sequences from angiosperms were aligned to identify conserved motifs and to derive a consensus from their extreme C-termini using WebLogo. There are two conserved motifs with Motif I (DxFFDDFP) being less well conserved than Motif II (DMFD). The derived consensus is strongly influenced by the overrepresented core eudicot group. The sizes of the letters represent the relative conservation, or the proportion of the sequences that has the same residue at the site.

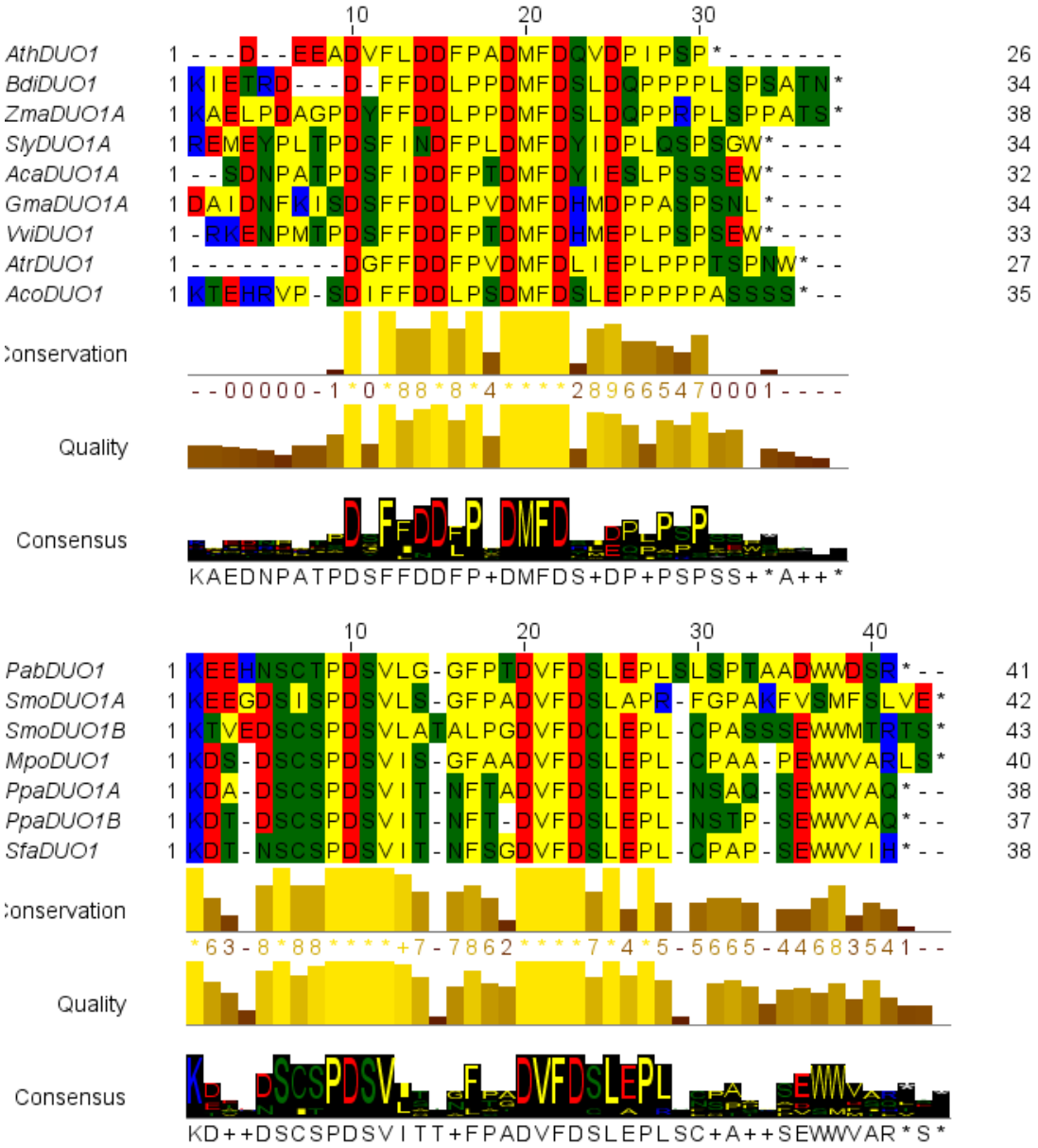


Figure 6. 2 Consensus of DUO1 C-termini for angiosperms and pre-angiosperms. Nine species representing major angiosperm clades were selected (basal angiosperm, basal eudicot, commelinid monocots, rosid, asterid). The same two previous motifs can be observed. All available pre-angiosperm DUO1 sequences within Embryophyta were aligned. The DUO1 sequences of *Selaginella moellendorffii* and *Sphagnum fallax* were not predicted by the source but through the method described in Chapter 3. The pre-angiosperms contain the same second motif (DVFD, instead of DMFD), yet showed no sign of the first motif as in angiosperms (DxFFDDFP). Instead, they have their own consensus (DSCSPDSV) slightly upstream of where the angiosperms have their first motif. Alignment and consensus sequences are displayed using Jalview. Ath = *Arabidopsis thaliana*, Bdi = *Brachypodium distachyon*, Zma = *Zea maize*, Sly = *Solanum lycopersicum*, Aca = *Aquilegia caerulea*, Gma = *Glycine max*, Vvi = *Vitis vinifera*, Atr = *Amborella trichopoda*, Aco = *Ananas comosus*, Pab = *Picea abies*, Smo = *Selaginella moellendorffii*, Mpo = *Marchantia polymorpha*, Ppa = *Physcomitrella patens*, Sfa = *Sphagnum fallax*.

Significant changes happened for the male germline cells during that period. Diploid stage became the dominant phase of the life cycle, and the sperms lost the ability to swim. None of these events caused any significant changes in the DUO1 sequence. This indicates that the rapid changes in protein length and C-terminal sequences were due to the evolution in the angiosperms. Thus, Motif I emerged after the divergence of gymnosperms and angiosperms and this might have resulted in the functional differences between the early and later evolved DUO1 proteins. Due to the lack of a 3D model, the only approach was to see if any significant secondary structure is present at the C-terminal region. The secondary structures of the same DUO1 homologs tested for hydrophobicity in Chapter 3 were predicted using SPIDER2 again (Figure 6. 3).

```
AtDUO1
SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLDDFPADMFDQVDPIPSP 297
SS  : 251 ----------------------------------------------- 297
rASA: 251 353555426546345536634675533322451345215456747 86 297
OsDUO1*
SEQ : 277 ELLPMVQSVPMIMPFFGMECAHDAVKHGAFDDLPPNMFDDAVDQPPPPPPPPP 330
SS  : 277 ---------------------------------H-H---------------- 330
rASA: 277 5434556645543343455244534643104513552266557667776777 77 330
LlDUO1
SEQ : 279 TVELPIFASSGSGQQGTPNDNSNSGEFDDFFNELPDMFDYWEDPPPG 325
SS  : 279 -------------------------H-HH------HH--------- 325
rASA: 279 6463554556655655445665545425412551342253257458 8 325
MpDUO1*
SEQ : 346 EKLPMLYSNDPSIQEPDVKDSDSCSPDSVISGFAADVFDSLEPLCPAAPEWWVA 399
SS  : 346 E---H-E----------------------HHHHHHHH--HHH----HHHHHH 399
rASA: 346 551443235466376664654761546332432135115314620531352023 399
SlDUO1A
SEQ : 270 RSSEIGVKREMAEIGVKREMEYPLTPDSFINDFPLDMFDYIDPLQSPSGW 319
SS  : 270 --------HHH-HH---------------------------------- 319
rASA: 270 556555465544533455536433335211441344113313446474 53 319
SlDUO1B
SEQ : 255 IDVLGHGFDNVEIPYVNRQIEKPLTPDSFIDDFPLDMFEHIEPLQSPSQW 304
SS  : 255 ----------------------------------HHH------------ 304
rASA: 255 334334535445244345536543434211331235114324527566 53 304
```

Figure 6. 3 Secondary structure prediction of the DUO1 homologs using SPIDER2. These homologs were also chosen for the functional tests in Chapter 4. Only 3 or more residues connected together that marked with "H" were considered to have a potential helix. Although no secondary structure was predicted for AtDUO1 at the C-terminus in Chapter 3, the MpDUO1 and SlDUO1B both have a potential helix at the DMFD motif (yellow boxes). Both of them had a low activity in the transient assay compared to the other homologs. * OsDUO1 ends at 343 aa and MpDUO1 ends at 402 aa. Blue numbers indicate low relative accessible surface area.

For some species, certain residues have high possibilities to form helices, but it requires 3 or more residues connected together to have a potential helix. That leaves only the MpDUO1 and SlDUO1B with a potential helix at the DMFD motif. Both of them had a low activity in the transient assay (Chapter 1 and 4) compared to the other homologs. Thus it is possible that these residues in this position need to sit on the surface of the structure thus being in a helix might inhibit with its function at the DUO1 C-terminus.

Since little information exists on the secondary and tertiary structure of the C-terminus other than what has been done in this work, the study of its specific function focused on the amino acid sequences.

Using WebLogo, the two conserved sequence motifs from angiosperm DUO1 sequences were highlighted as before: DxFFDDFP and DMFDxxE/DP (Figure 6. 4). These two sequences of the motifs are similar to the 9-amino-acid Transactivation Domain (9aaTAD) sequences. Two highest score patterns predicted by the online "Nine Amino Acids Transactivation Domain 9aaTAD Prediction Tool" (http://www.med.muni.cz/9aaTAD/index.php) on the AtDUO1 C-terminus contain 7 out of 9 identical residues to known 9aaTAD members (Piskacek et al., 2015).

Key features of 9aaTADs are their acidity (Hope and Struhl, 1986, Gill and Ptashne, 1987), hydrophobicity (Drysdale et al., 1995), or both (Regier et al., 1993, Sainz et al., 1997, Sullivan et al., 1998). There is evidence that even with similar chemical properties, not every residue contributes equally and that substitution of a single amino acid in the 9aaTAD can significantly reduce transcriptional activity (Sainz et al., 1997).

Due to the diversity of 9aaTAD sequences, it was difficult to predict which residues might be critical within the C-terminal region of AtDUO1, but there were four potential 9aaTAD sequences based on the positions of the second consensus.

1. 2-10 (xxxDxFFDD)
2. 5-13 (DxFFDDFPx)
3. 9-17 (DDFPxDMFD)
4. 13-21 (DMFDxxEPx)

Although two 9aaTADs present in a single region is not unheard of (e.g. p53), it is very common for the 9aaTADs to be surrounded by similarly acidic and hydrophobic residues. So it is very likely that the true 9aaTAD of DUO1, if exist, locates at one of

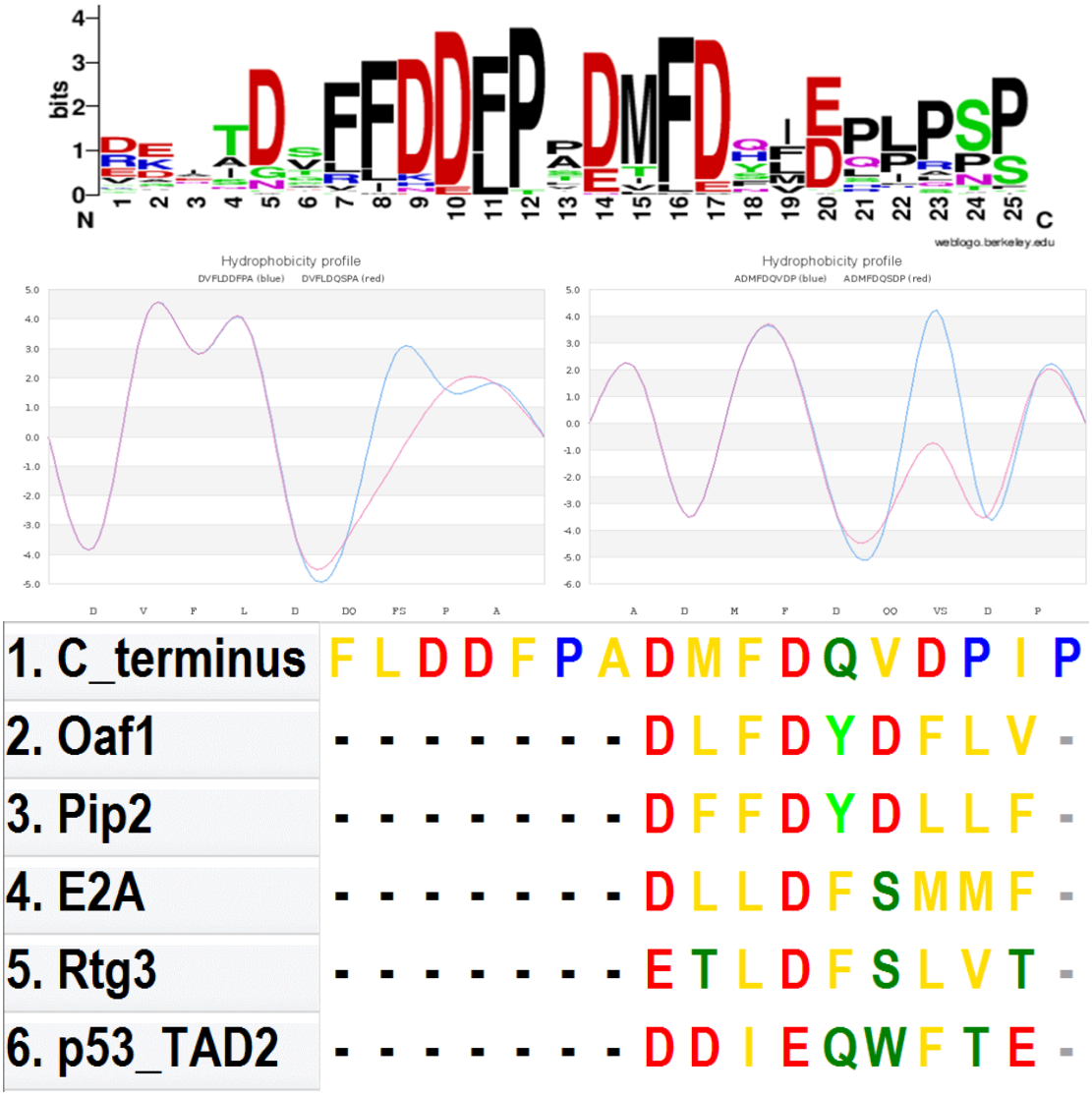


Figure 6. 4 Angiosperm DUO1 C-termini consensus have features like 9aaTADs. There are two motifs within the angiosperm DUO1 C-termini. Their hydrophobicity profiles (blue curves) fits well with some of the known 9-amino-acid transactivation domain (9aaTAD) members (red curves). Manual alignments against known 9aaTADs show that there are high similarities between the known 9aaTADs and AtDUO1 C-terminus.

the two conserved motifs, and the other was evolved to fine-tune its function. Given this hypothesis, it is more likely the case that the core sequence in all land plant DUO1 Motif II is the 9aaTAD and Motif I in different clades evolved to adapt its functions.

Nevertheless, some known 9aaTAD sequences were manually aligned with DUO1 Motif I (DxFFDDFP) sequences in the C-terminal conserved region. Some of the same sequences could also be aligned to DUO1 Motif II (DMFDxxE/DP) sequences. Regardless of the true beginnings and endings of the two motifs, the amino acid substitutions within the most conserved residues of the two motifs (8-11 FLDD and 14-

17 DMFD) were predicted to affect the properties and result in reduced transcriptional activation by DUO1.

Identifying these key residues was crucial for designing experiments to confirm the potential TAD property of the AtDUO1 C-terminus. The importance of key residues is testable and has been tested in many TAD studies.

6.3 Deletion and site mutation of the extreme C-terminal end

To test the functional importance of these conserved regions, and identify the key residues in the C-terminus that are responsible for these functions, three AtDUO1 variants and three MpDUO1 variants were constructed.

For AtDUO1, these were FDQV → FDSL (AtDUO1FDSL), which served as a control that expected to have no change in activation ability; FLDD → FLSG (AtDUO1FLSG), which interfered with the first motif and expected to have a moderate impact on the transcriptional activity; and DMFD → KMFK (AtDUO1KMFK) which disrupted the core of the second motif and should have the strongest effect. The secondary structures were predicted using SPIDER2 (Figure 6. 5). The latter two also showed a helix at the extreme C-terminal end.

These constructs were tested using the dual-luciferase assay in tobacco leaves (Figure 6. 6). The *AtHTR10* promoter was chosen as the reporter for these effectors. The results confirmed the prediction to a high degree. The activity of the DMFD → KMFK variant was the most severely reduced. The FLDD → FLSG variant which reduced the acidity of the C-terminus had a moderate effect on activity, and the FDQV → FDSL variant, with two substitutions at a non-conserved site, showed little change in activation.

A parallel change of the DVFD → KVFK, and two reciprocal changes to the MpDUO1 C-terminus were made by substitution in the AtDUO1 sequences (i.e. VISG → VIDD & FDSL → FDQV). For the *AtHTR10* promoter activity, DVFD → KVFK reduced the low level activation of the MpDUO1 further as expected. VISG → VIDD was predicted to increase activation, but no statistically significant increase was observed. FDSL → FDQV showed a slight decrease of activation, although no significant difference was expected.

A GFP protein was C-terminally fused to all six constructs. They were tested in the same assays as a repeat. The same parts of the leaves injected with the *Agrobacteria* were then placed under a microscope to exam the GFP signal. The activation levels for these constructs had the same trend as their no-GFP counterparts. The GFP signals within each group were also comparable between constructs (see Appendix).

The effect caused by residue changes in AtDUO1 C-terminus, in the case of FLDD → FLSG could not be reversed in MpDUO1. It made distinguishing the molecular mechanism of this activation difference much more difficult. The change of polarity or electric charge could be directly affecting the target interaction, or may have disrupted the local structure of the motifs as predicted by the secondary structure models.

```
AtDUO1
SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLDDFPADMFDQVDPIPSP 297
SS  : 251 ---------------------------------------------- 297
rASA: 251 353555426546345536634675533322451345215456747 86 297
AtDUO1FDSL
SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLDDFPADMFDSLDPIPSP 297
SS  : 251 -----------------------------------H---------- 297
rASA: 251 353545426546345536535565533221441345215556737 86 297
AtDUO1FLSG
SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLSGFPADMFDQVDPIPSP 297
SS  : 251 ---------------------------H-----HHH---------- 297
rASA: 251 353545436546345535524565423211441336114536637 76 297
AtDUO1KMFK
SEQ : 251 FALPQPFFEPSPVPRRCRHVSKDEEADVFLDDFPAKMFKQVDPIPSP 297
SS  : 251 -----------------------------------HHH---------- 297
rASA: 251 354555436556346636634565534312441345216546737 86 297
```

Figure 6. 5 Secondary structure predictions for the site-directed mutation variants. Similar to the MpDUO1 and SlDUO1B having a potential helix at the DMFD motif, FLSG and KMFK both showed a potential helix, and had a lower activity in the transient assay (Figure 6. 6). This agrees with the possibility that a helix in this position might be detrimental for DUO1 C-terminus activation ability. Yellow/cyan boxes = rASA changes.

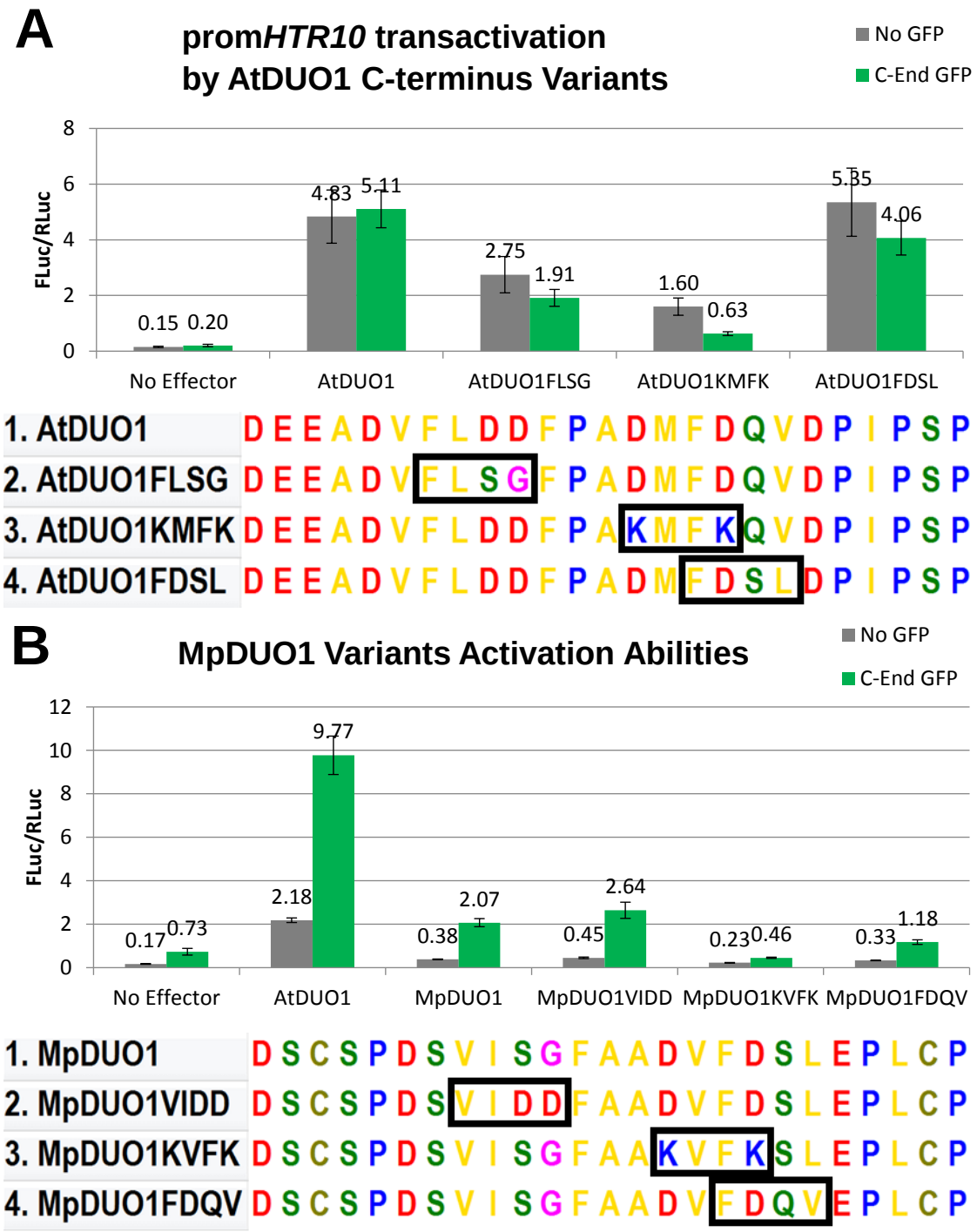


Figure 6. 6 Effects of AtDUO1 and MpDUO1 site-directed mutagenesis on the *HTR10* promoter transactivation.

Reducing the acidities in the conserved AtDUO1 C-terminal motifs decreased the activity of two mutants (DMFD → KMFK & FLDD → FLSG), while the control mutation (FDQV → FDSL) showed no significant difference from wild type DUO1. The disruption of the KMFK motif yielded the most dramatic effect, and the change of the acidity in the FLDD motif also significantly reduced the activation ability. Similarly, the change to the core DVFD motif to KVFK decreased the activity of the MpDUO1. However, the control mutation had a slight decrease compared to its wild type version, when expected to be the same. The increase of the acidity in Motif I FLSG → FLDD did not produce a statistically significant increase. The tests were repeated after fusing a GFP protein at the C-terminal end of each variant.

6.4 C-terminal end is not an independent transactivation domain

In Chapter 4, it has been discussed that the MpDUO1 cannot efficiently activate the AtDUO1 targets, but MpChimera which changed the MpDUO1 C-terminal region with the AtDUO1 successfully raised its activation abilities. A key part responsible for the AtDUO1 activation abilities lies within in the last 25 amino acids AtDUO1 C-terminus Transactivation Domain (AtDUO1 TAD). Yet the key residue switches in the MpDUO1 TAD failed to raise the activation abilities for MpDUO1. This indicates that the rest of the C-terminal region, referred to as the linking region in Chapter 3, also plays a key role in the activation abilities of DUO1.

Five constructs were made to test this linking region (Figure 6. 7):

1. MpDUO1-AtAD attached the AtDUO1 TAD at the end of the MpDUO1
2. MpDUO1ΔC1-AtAD switched the MpDUO1 TAD with the AtDUO1 TAD
3. MpDUO1ΔC2-AtAD deleted the 100 amino acids that are next to the TAD (C2) from the MpDUO1ΔC1-AtAD
4. MpDUO1ΔC3-AtAD deleted the whole C-terminal region except for the 100 100 amino acids that are immediately after the MYB domain (C4), then added the whole AtDUO1 C-terminal region
5. MpDUO1ΔC4-AtAD deleted the 100 amino acids that are immediately after the MYB domain (C4) from the MpDUO1ΔC1-AtAD

The first two constructs added the AtDUO1 TAD onto MpDUO1 directly or replaced the MpDUO1 TAD. They were designed to test the functional independency of the AtDUO1 TAD, which evidence from the MpDUO1 TAD mutant variants suggested to be not the case. In the transient assays, neither of these constructs showed a significant increase of target *HTR10* promoter activation compared to the MpDUO1. Interestingly, while the activation level of the MpDUO1ΔC1-AtAD showed no statistical difference at all, MpDUO1-AtAD had a significant decrease in comparison to MpDUO1.

There are two simple explanations for this result. The first one is that the activator abilities of the AtDUO1 TAD require a larger region than the last 25 amino acids. The second one is the potential constraint for the linking region length in the angiosperm DUO1 (Chapter 3). To test if AtDUO1 TAD needs a larger region, MpDUO1ΔC2-AtAD and MpDUO1ΔC4-AtAD were constructs based on MpDUO1ΔC1-AtAD but kept the

linking region length consistent with the AtDUO1 and MpChimera. Two separate regions (C2 and C4) were deleted to avoid the possible influence of any potential inhibiting sequence. Neither constructs were able to increase the activation level. This means the AtDUO1 TAD requires certain upstream sequences to function properly.

To test if the linking region length is restricted, MpDUO1ΔC3-AtAD extended a fragment from MpDUO1 after the MYB domain (C4) from the MpChimera. This made the elongated MpChimera the same length as the MpDUO1. Since the MpChimera had activation abilities both *in vivo* and *in planta*, the whole AtDUO1 C-terminal region definitely contained the proper sequences required for transactivation. This construct did not show an increased activation ability in the transient assay, either.

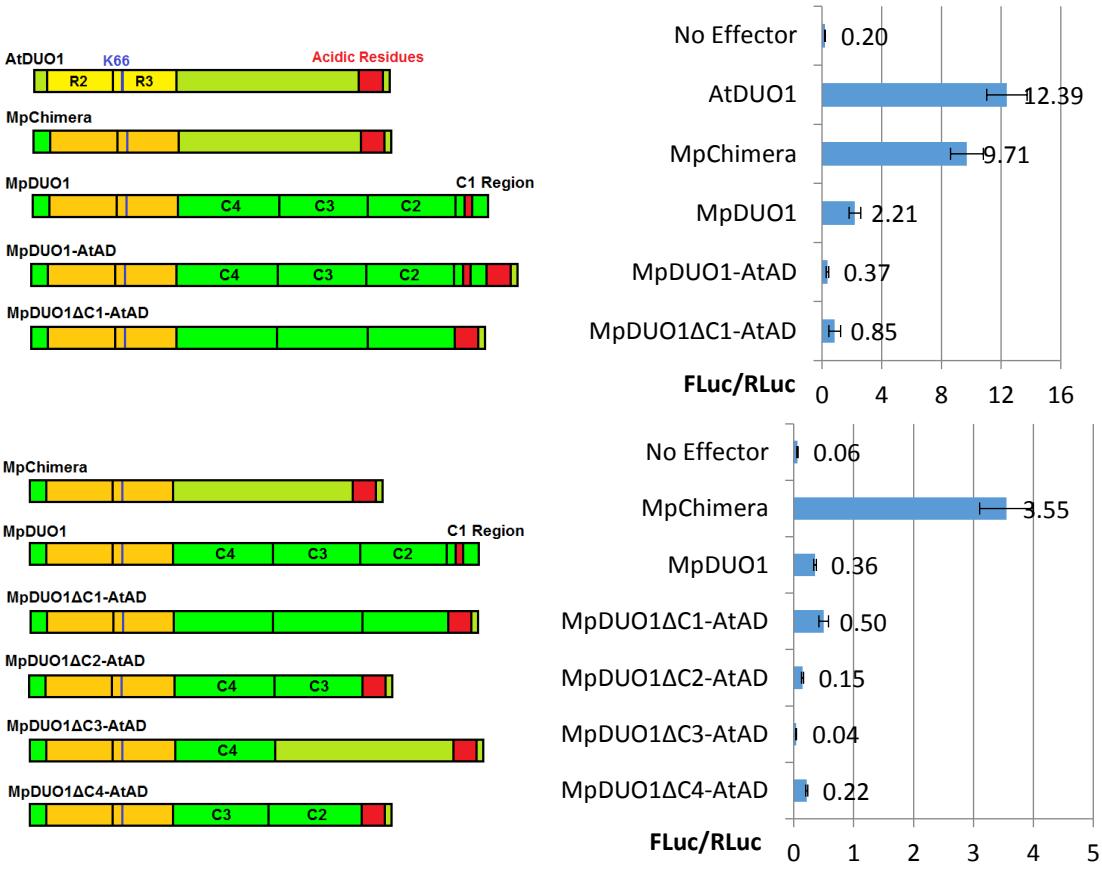


Figure 6. 7 A series of variants using AtDUO1 and MpDUO1 fragment combinations. MpDUO1 cannot activate the AtDUO1 targets efficiently, while MpChimera is confirmed to be functional (Chapter 4). Although the C-terminal end region (C1) contains key changes that are related to high level of target activation, the rest of the sequences at the C-terminal region may also have a strong influence. Five constructs were built to determine whether the linking region length or the adjacent sequence next to the extreme AtDUO1 C-terminus (25 aa) is essential for its transactivation abilities. None of these constructs had an increased target *HTR10* promoter activation level (blue bars) in the transient assay. Yellow = MYB domain; Red = C=terminal motifs.

Notably, there were some differences in the activation abilities of these constructs. The switch of MpDUO1 TAD to the AtDUO1 TAD did not alter the activation level. This agreed with the results of previous experiments where the increase of acidity did not affect the performance of the MpDUO1 TAD. The shortening of the linking region slightly but significantly reduced the activation ability, while lengthen of the linking region to the known activator MpChimera completely dismantled its activation function.

6.5 *A. thaliana* DUO1 has activation functions across kingdoms

Dr Borg has previously shown that for AtDUO1, the MYB domain alone did not have any activation capacity, while the removal of the TAD 25 amino acids (AtDUO1ΔC3) reduced the activation ability to an extremely low level (Borg, 2011). Due to the lack of any visible sequence conservation, after the functional analysis from OsDUO1 and SlDUO1A from Dr Sari (Chapter 1 and 4), it was first speculated that any unstructured region would be sufficient to form a linking region to support the function of the DUO1 TAD. The functions of the DUO1 domains under this hypothesis can be summarised as Figure 6. 8 Model 1, in which the MYB domain is for DNA binding, the linking region and the TAD are a weak and strong activator, respectively.

However, the results from the MpDUO1ΔC constructs indicate that the linking region plays a vital role in the activation function of DUO1. The activation ability of the DUO1 TAD seems to be strictly limited by the linking region. In light of this new evidence, a second model was proposed in Model 2. While the linking region is still being considered an activator, the TAD is instead being viewed as an amplifier, not an independent activator. In other words, the revised DUO1 TAD now includes the extreme C-terminal end and a part of or the whole linking region.

Under this model, the mutations at the AtDUO1 C-terminal end altered the amplifying quality of the module. MpDUO1 C-terminus had already boosted the level to the maximum capacity allowed by its linking region, the changes (MpDUO1VIDD & MpDUO1ΔC1) could not further raise it. The shortening of the MpDUO1 length (MpDUO1ΔC2/4) lower the linking region activation ability as a result of disruption. For MpDUO1ΔC3, the linking region was a mix of MpDUO1 and AtDUO1 sequence and had no activation ability.

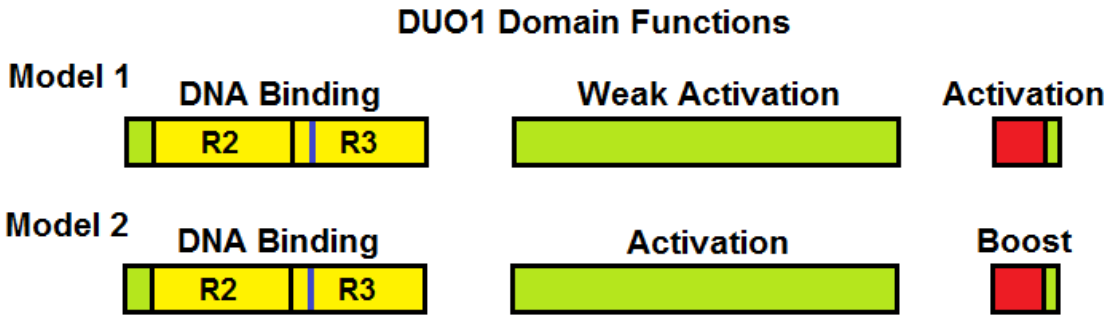


Figure 6. 8 Two different models for the functions of different DUO1 domains.
The MYB domain has been established as the DNA binding domain. In Model 1, the linking region and the extreme C-terminus are two separate activation domains, the latter being a much more potent one. In Model 2, the two parts are one integrated module, the linking region is the activator and the extreme C-terminus enhance the activation level.

It also explains how algal DUO1 homologs can function as potential activators even though they lack any recognisable C-terminal motif. At its origin DUO1 might not have the need to boost its activation ability. It is important to recognise that this model only tries to tell which role each domain plays in the activation function, but does not explain the mechanisms of the activation.

If this model is correct, then somewhere in the linking region lies the sequence responsible for the activation function of DUO1. To further investigate the role the linking region plays in activation, the test of AtDUO1ΔC3 was repeated in the transient assays with two more constructs, AtDUO1ΔC1 and AtDUO1ΔC2, which contained further C-terminal deletions into the linking region (Figure 6. 9).

The result of AtDUO1ΔC3 agreed with the data from Dr Borg, which shows a huge drop in its target *HTR10* promoter activation level compared to the full length AtDUO1. DUO1ΔC2 (233 aa) was 39 amino acids shorter than ΔC3 (272 aa), and ΔC1 (180 aa) was 53 shorter than ΔC2. No significant difference in their target activation levels between them was detected, but ΔC1 and ΔC2 only had about half of the level compared to ΔC3. This significantly reduced activation level demonstrates that the linking region does have sequences that are important for target activation, despite the lack of notable conservation. Interestingly, ΔC1 and ΔC2 still had some activation abilities left, suggesting the linking region might contain two different parts for activation. One of them is located in the 39 amino acids deleted from ΔC3 in ΔC2, while the other is between the MYB domain and the end of ΔC1.

It is rather curious how DUO1 can be replaceable with a homolog (e.g. OsDUO1) containing no obvious sequence conservation in the linking region, which may play a role so essential for activation and have a complex structure consists of different parts. This makes the conserved C-terminus TAD even more interesting.

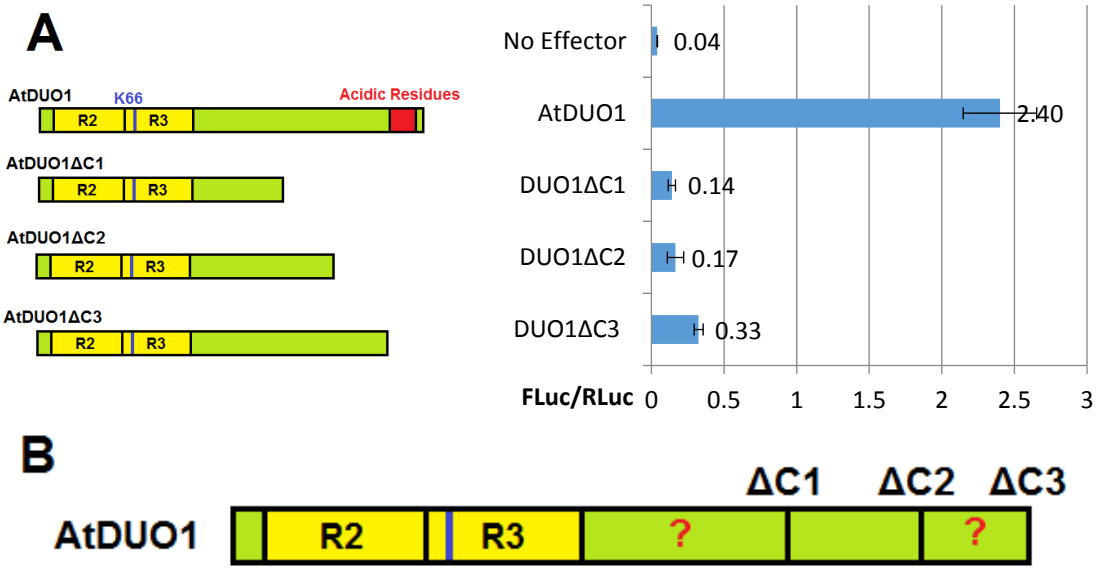


Figure 6. 9 All three C-terminal deletion variants have activation abilities.
A. The results of the ΔC1 and ΔC2 were not significantly different from each other. They had a significantly lower activation ability than DUO1ΔC3, which was less than 20 % of native full length DUO1. This demonstrated that the linking region of AtDUO1 has some activation ability. B. The most likely scenario is that the activation ability is contributed by two possibly separate parts (red question marks).

The DUO1 extreme C-terminal region is essential for the high level of activation of its targets. So far several key residues have been identified as important for its function, which possess charactistics of a transactivation domain. VP16 has been shown to be a potent transcriptonal activator in mammalian cells (Sadowski et al., 1988, Uesugi et al., 1997), as well as in plant cells (Storgaard et al., 2002, Silveira et al., 2007). It was shown that it might function in a similar mechanism through interaction with the mediator complex, both in human cells and in *Arabidopsis thaliana* (Aguilar et al., 2014). Considering the similarity in sequences, it is possible that the DUO1 C-terminus is an interchangeable member of a large family of transactivation domains that includes VP16.

One of the most famous example is GAL4 from yeast, which is known to function in animal and plant cells (Kakidani and Ptashne, 1988, Ma et al., 1988). Similarly, plant

TADs such as Dof1 has been confirmed to function in animal and yeast cells (Yanagisawa, 2001). Transcription factors ST1 from the plant *Nicotiana tabacum* and MSN1 from the yeast *Saccharomyces cerevisiae* have also been reported to be able to influence one another in both systems (Kim et al., 2006). More experiments on plant TADs have previously been done in yeast with VP16 as a control (Tiwari et al., 2012, Li et al., 2013a).

To investigate this hypothesis, a hybrid DUO1 molecule was created of which the AtDUO1 TAD was replaced with the VP16 TAD. It was decided that the mammalian cell system previously assayed with a plant TAD (Dreb2a) and VP16 TAD (Aguilar et al., 2014), along with the usual plant system should be sufficient for the test to get a definite answer.

The molecules were tested for complementation *in planta*. DUO1ΔC3 was capable of fully rescuing the cell division defect phenotype, but not activating the cell differentiation marker *HTR10*. It failed to transmit through the male when crossed with *ms1* plant, either (Borg, 2011). When the VP16 domain was substituted for the AtDUO1 TAD (DUO1ΔC3VP16), it was able to fully rescue the cell division defect and activate the cell differentiation marker (Figure 6. 10). It was also fully transmitted as shown in Table 6. 1.

Table 6. 1 AtDUO1ΔC3VP16 rescues the male transmission of *duo1-1*

					χ^2 Significance	
Rescuing Transgene	Lines	PPT^R	PPT^S	PPT^R:PPT^S	2:1	1:1
promDUO1:AtDUO1-mCherry	2	132	65	2.0:1.0	ns	***
promDUO1:AtDUO1ΔC3VP16-mCherry	3	467	231	2.0:1.0	ns	***

The promDUO1:AtDUO1ΔC3VP16-mCherry was introduced into *duo1-1*$^{+/-}$ plants, and seedlings from single locus T1 lines scored for PPT resistance; PPT^R = resistant seedlings counted, PPT^S = sensitive seedlings counted. χ^2 analysis was used to test for significant deviation from the expected ratio of 2:1 if there was a full transmission for the transgene, or 1:1 if there was no transmission; the AtDUO1 control data form Sari, 2015. ***P < 0.001. ns, not significant.

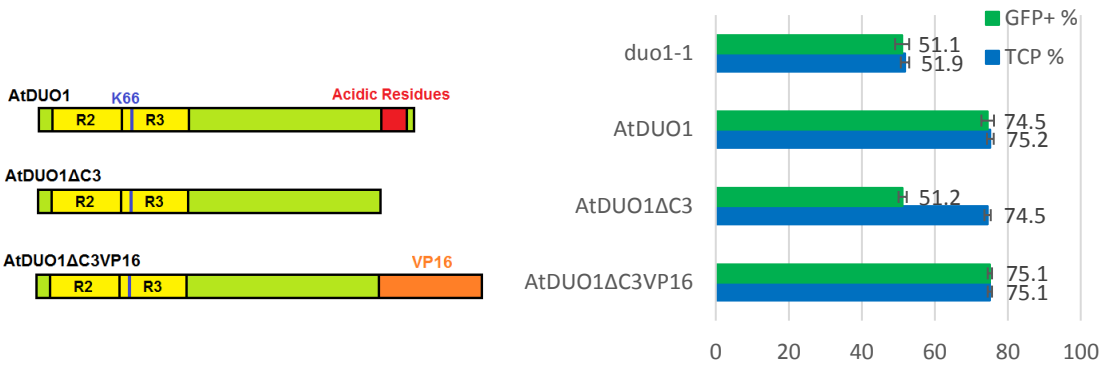


Figure 6. 10 VP16 can substitute the AtDUO1 TAD activation abilities *in planta*.
The DUO1ΔC3 can rescue cell division defect but not cell differentiation. By attaching VP16 onto the C-terminus it was able to fully rescue both pathways *in planta*. Error bars represent standard error, n = 3.

To get more quantitative data, these molecules were also tested in the transient expression assays in two different cell systems for the *Arabidopsis thaliana* promoters activation levels (Figure 6. 11).

In the transient assays (*Nicotiana tabacum* leaf cell), compared to full length AtDUO1, DUO1ΔC3 had less than 15 % of the activation level. After being attached with VP16, it was able to significantly boost the activation ability, with 60 % level of the wild-type for two promoters indicating the cell division and differentiation pathways (see Appendix for promDAZ1 activation).

In the mammalian two-hybrid assays (*Homo sapiens* HEK 293T cell), all three molecules activated the target *HTR10* promoter, demonstrating the cross-kingdom activation ability of DUO1 TAD as predicted. However, full length AtDUO1 had only double the activation level of DUO1ΔC3, a much lower ratio compared to the plant leaf cell system. Interestingly, the level of DUO1ΔC3VP16 was more than five times higher than the AtDUO1 level. This showed a sharp contrast and it is clear now that the activation abilities of the two TADs varies widely in different types of eukaryotic cells.

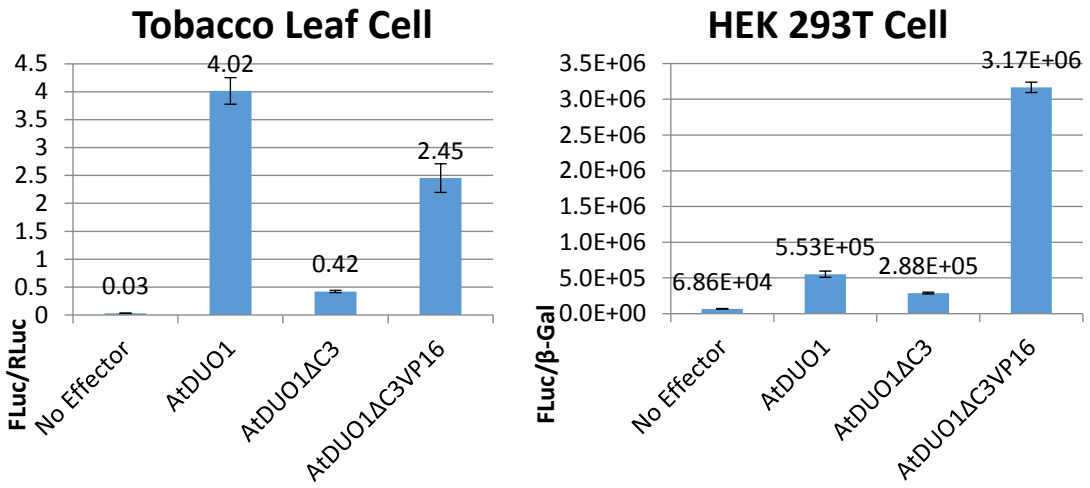


Figure 6. 11 VP16 and AtDUO1 TAD both have activation function across kingdoms. Compared to AtDUO1, DUO1ΔC3 had less than 15 % of the target *HTR10* promoter activation level in tobacco leaf cells. DUO1ΔC3VP16 was more than 5 times higher, about 60 % of the AtDUO1 level. The VP16 activity surpassed AtDUO1 in human cells, more than 5 times higher. Interestingly, the DUO1ΔC3 can also act as an activator in both systems, although to a much lower degree.

6.6 Discussion

The AtDUO1 C-terminus conservation provided a great insight into the conservation of the DUO1 function. Many DUO1 homologs in species with multiple copies have lost its C-terminal TAD (e.g. *Zea mays*, *Glycine max*, *etc.*). There have been some curious occasions. For example, potato (*Solanum tuberosum*) has a single copy with mutated DMFK. This is possibly due to the existence of another functional copy that was not found since only a homozygous doubled-monoploid potato clone was sequenced (Potato Genome Sequencing et al., 2011). Alternatively, this could be indicating reduced reproductive ability in favour of vegetative growth (MalkinaPykh, 1996).

DUO1 homologs exist in green algae and possibly function as an activator using the sequences that are not conserved in any clade. The C-terminal motifs likely emerged in land plants as an activation domain to accommodate the changes in reproductive gene network. The sudden shifts in sequence conservation and activation ability from the gymnosperms to angiosperms remain unexplained. In angiosperms, genes like *GCS1* and *GEX2* have been recruited into the DUO1 regulation network, consequently DUO1 might require a higher activation ability to synchronise their expressions. Angiosperm DUO1 proteins achieved that by developing a highly acidic transactivation domain. For all intents and purposes, the mechanisms of the TADs are yet to be understood.

In eukaryotic cells, almost all transcription of the RNA pol II promoter requires a mediator complex (Takagi and Kornberg, 2006). For instance, several known transcription factors of the 9aaTAD family, including Gal4, Oaf1 and Pdr1, have been reported to rely on the transcriptional mediator MED15 (Piskacek et al., 2016).

To look for a mediator candidate for DUO1, it is useful to look at its TAD replacement VP16. The key sequence for the VP16 activation is DFDLDMLGD (Sandholzer et al., 2007, Aguilar et al., 2014). VP16 was shown to bind both human and *A. thaliana* MED25ACID (ACtivation Interaction Domain) (Milbradt et al., 2011, Vojnic et al., 2011, Aguilar et al., 2014). AtMED25 is known to be involved in different stress response pathways (Elfving et al., 2011). The Dreb2a from *A. thaliana* can also interact with both MED25ACIDs, it contains a motif DMFDVDELL (Aguilar et al., 2014). The key motifs of these three proteins are highly similar (Figure 6. 12). The Dreb2a TAD even contains the sequence of DMFD. Therefore, AtMED25 is considered the primary candidate for interacting with DUO1.

Figure 6. 12 Alignment of the TADs from AtDUO1, AtDreb2a, and VP16.
All three TADs have a similar acidic/hydrophobic composition. AtDUO1 and Dreb2a TADs share the DMFD sequence, while VP16 keys residues include the sequence DDFD. The black box indicates the core TAD sequences.

A Glutathion S-transferase (GST) pull-down assay was used to identify the MED25 for interacting with AtDreb2a and VP16 TAD (Aguilar et al., 2014). A similar approach could be devised to confirm the mediator complex involved in the activation by AtDUO1 TAD.

Erikina and Erkine (2016) introduced an interesting idea for the possible mechanism of the TADs, namely nucleosome distortion. The concept of conformational rearrangement for "induced-fit" has been established in many instances, like target DNA recognition and binding (Velmurugu, 2016), DNA repair (Ghodke et al., 2014), homologous recombination (Rambo et al., 2010, Savir and Tlusty, 2010), tRNA decoding (Savir and Tlusty, 2013), and enzyme-substrate interaction (Savir and Tlusty, 2007).

The activation ability of the DUO1ΔC3 can be explained by the "intrinsic structural disorder" for nucleosome distortion. In both plant and mammalian cells, this low affinity interaction allowed general transcription factors to move in. VP16 also contains an "unstructured region" for it to stabilize on the DNA and interact with HCF-1 and Oct-1 (Liu et al., 1999, Wysocka and Herr, 2003). However, it does not explain how DUO1ΔC3 managed to recruit the co-activator complexes, as it lacks any conceivable conservation. Maybe a much better explanation is, it does not.

The mediators do not seem to interact with DNA directly, but instead is a part of the preinitiation complexes on the enhancer that pass on the signals from the transcription factor (activator or repressor) to the RNA pol II (Kornberg, 2007). As for now, there is no evidence of interaction between the mediator complexes and other general transcription factors (GTF) either. In other words, it is safe to considered that the mediators are recruited by specific transcription factors only. In the experiment, the removal of the TAD for AtDUO1 dramatically reduced the activation ability. This indicates that the TAD, or at least this 9aaTAD type, is able to induce high level of activation by the conventional method of recruiting mediators. Any disruption of the interaction between the TAD and the mediator complex would be an equivalent of an inhibition of the enhancer.

Savir and Tlusty (2007) in his research of enzyme catalysis proposed deformations upon binding as a potential conformational proofreading mechanism for optimal specificity. The idea is that there is a conformational mismatch between the ligand and its main target in their intial encounter. A similar matching mechanism induced by conformational changes has also been reported recently for the human cannabinoid receptor CB1, which allows CB1 to bind to different types of molecules (Hua et al., 2016, Shao et al., 2016, Hua et al., 2017). This could explain why so many TADs look similar and function across kingdoms, as the intial interactions for recruiting other transcription co-activators like mediators can happen easily with very limited similarity. However, the following deformation proofreading plays a large role in the level of activation. This is only possible thanks to the flexibility provided by the intrisically disordered regions in the mediator subunits (Allen and Taatjes, 2015). Therefore, the AtDUO1 TAD has a better specificity for the plant (i.e. *A. thaliana* and *N. tabacum*) MED25, while TADs from other proteins like VP16 have reduced activation levels. In mammlian cells, VP16 TAD has the optimal specificity, while AtDUO1 TAD failed to pass the secondary test.

This hypothesis actually points away from the differences at the TADs regarding the activation level differences between different plant DUO1 proteins in the transient assays. The affinity of the NtMED25 to the SlDUO1A TAD is surely equal to or even higher than to the AtDUO1. Instead, the differences are derived from other regions. One of such a difference is the target sequences recognition and binding abilities of the MYB domains, which is demonstrated by the PpChimera and MpChimera experiments that even without a changed target sequence, the activation level can be reduced to half. Another one may lie in the linking region, which carries out the nucleosome distortion function. Thus only the AtDUO1 protein matches the AtDUO1 target promoter perfectly to allow these two interactions happen to the fullest extent.

This also explains why the efforts to increase the activation ability of MpDUO1 did not prove successful. The linking region of MpDUO1 could not deform the AtDUO1 promoter properly, but it does not need to do so to carry out its function in *Marchantia polymorpha*. The MpDUO1 TAD still has its activity (demonstrated by the reduced level in the DVFD → KVFK construct), and is optimised to the level which its linking region allows the binding of the GTFs. Therefore, as long as the linking region consists of the MpDUO1 sequence, the higher affinity between the TAD and the mediator cannot override the GTFs binding defect in the tests. In the case of MpDUO1ΔC3-AtAD, an additional 100 amino acids was attached after the MYB domain of MpChimera, resulted in a combination of MpDUO1 and AtDUO1 linking region sequences. This could lead to a reduced flexibility and thus completely disrupting the nucleosome distortion ability.

However, it is important to note that this hypothesis is based on the "Occam's Razor" principle with certain assumptions. Most importantly, it assumes that better affinity due to co-evolution would lead to higher activation level, which is not necessarily true. It is possible for a lower activity to be an intrinsic property being selected for in certain environment. The hypothesis would not be affected should this be the case, since the hypothesis explains the mechanism (how) while the evolution explains the reason (why). The related but more testable assumption is that for a transcription factor controlling a network involving many sets of genes, most angiosperm groups should have comparable levels of DUO1 targets expression (see Gene Dosage Hypothesis in Chapter 1). This in turn explains the experiment results as that DUO1 and its targets within the same species should have higher affinity than compared to the cross species interaction. This is also verifiable using the promoters from another species (e.g. *Solanum lycoperiscum*) and

then tested with AtDUO1 and SlDUO1A. The hypothesis predicts a reverse in the activation level differences. Another assumption testable is that the different DUO1 proteins are expressed at the same level. This can be tested by attaching a GFP at the C-terminus and then measure the florescent levels as described in this thesis. The overall hypothesis can be heuristically modified based on the results of these tests.

Chapter 7 Conclusions & Perspectives

7.1 Summary

In *Arabidopsis thaliana*, DUO1 is required for sperm cell cycle progression and cell differentiation. The expression of DUO1 is male germline-specific, subsequently its direct target genes such as *DAZ1/2*, *HTR10*, *GCS1*, *GEX2*, etc. are also male germline-specific (Brownfield et al., 2009a). DUO1 and members of this network such as DAZ1 have homologs in all land plants, indicating an entirely or partially conserved regulating network for the plant male germline (Borg et al., 2011).

Understanding the evolution and conservation of the DUO1 function was the main aim of this thesis. In Chapter 3, sequence analysis of the DUO1 homologs collected revealed the relationship between the DUO1 and the GAMYB protein sister clades. Their divergence in the green algae Charophyta marked the emergence of the early DUO1 proteins. As both clades are related to male reproduction in every known case, it is reasonable to suspect that the ancestral R2R3 MYB protein had a similar role in algae. In contrast to DUO1, GAMYB family members are not male germline proteins. Evidence suggests that the function of DUO1 is confined to male gamete development and linked to the sexual reproduction in the whole Streptophyta clade. This high specificity of function may have been responsible for the lack of proliferation in the DUO1 clade (Dubos et al., 2010).

The protein sequence and tertiary structure differences in the MYB domain between DUO1, GAMYB, and ancestral R2R3 MYB proteins provided evidence of key changes in the region resulted in the divergence. It also suggested a secondary shift in target DNA recognition residues in the surviving Embryophyta, which has been fixed and maintained. The two events happened ca. 900 mya and 500 mya, respectively (Clarke et al., 2011, Magallon et al., 2013). The sequence of the C-terminal part indicated a third shift in the angiosperms, which dates to ca. 150 mya (Magallon et al., 2015, Foster et al., 2017).

Functional tests agreed with the significance of the last two shifts. It was demonstrated that the MYB domain of all Embryophyta could recognise and bind to the same target DNA sequence. While all functional angiosperm DUO1 homologs were able to show a

comparable level of activation function to AtDUO1, bryophyte DUO1 homologs tested were unable to do so due to the C-terminal region differences. Interestingly, the miR159 binding site in the MYB domain was acquired after the establishment of the core angiosperm group, even though the structural change happened right after the emergence of the early DUO1. It is clear that these two events happened independently, and the recruitment of the miR159 regulation at that site was a coincident.

The extreme C-terminal end has a disproportionate importance on the activation ability of DUO1. Chapter 6 was dedicated to the understanding of the activation mechanism of DUO1 and how the changes from pre-angiosperms to angiosperms had led to the differences in the target activation. After the identification of some key residues, it was proposed that the DUO1 is part of the nine-amino-acid transactivation domain (9aaTAD) family (Piskacek et al., 2016). Resemblance to the known case of AtDreb2a and VP16 TAD and further functional tests suggested that DUO1 TAD might use the same mechanism as these two, which involves other general transcription activators and the MED25-induced mediator complex (Aguilar et al., 2014).

7.2 Streptophyta offers a case study for sexual reproduction

There is a notorious high cost on sexual reproduction, or to be more precisely, anisogamy. This is the simple mathematics of a 50 % reduction on offspring that reproduce. Yet most branches of life have developed sexual reproduction independently. The most plausible answer available so far is the benefits of recombination (Burt, 2000). In addition to the increase of genetic variation, it also offers DNA repair and complementation. A faster rate of evolution gives a population more chances to take different niches. DNA repair allows an individual that suffers a faulty genome to survive, but this does not offer an advantage over asexual reproduction since in both case one functional and one faulty genome would produce one viable individual. Complementation on the other hand, can protect deleterious mutations from being eliminated that would provide fitness advantage in a changed environment, such as the famous case of the sickle-cell disease (Serjeant, 2010). This resistance against parasites using sexual reproduction has been considered as a key adaptation (Hamilton et al., 1990, Morran et al., 2011). Opposite of this view that sexual reproduction drives more genetic variation, some has argued about its ability to clean out the harmful mutations within the population (Gorelick and Heng, 2011).

However, this again offers no advantage over asexual reproduction as deleterious mutations would be eliminated even faster, and therefore will not be further discussed here. The hypothesis that sexual reproduction creates more genetic variation and consequently expedite evolution agrees with the observation that sexual reproductive genes seem to evolve at a higher rate (Wilburn and Swanson, 2015). Although this hypothesis can be challenging to test empirically, advances in computer simulation are offering us more opportunities to explore this question (Stauffer et al., 2001).

A more bizarre but nevertheless interesting question arise as to why independently evolved sexual reproduction systems are all exclusively binary (Czaran and Hoekstra, 2004). To be specific, this discussion excludes a multiple mating type system that does not allow all gametes to randomly pair up, like certain harvest ants (Parker, 2004). The answer may lie in simple mathematics again, as shown in the case of Streptophyta.

If all mating types are needed, a three-part system would sacrifice 66 % of the resources. Like in the animal kingdom, sexual reproduction has been reversed in several branches of Streptophyta like Zygnematales in this thesis, with its sexual reproduction genes degenerated. Exclusively self-pollinating species have also essentially traded genetic variation for reproduction success rate. Similar to the cases of many vertebrates but much more prevalently (Booth et al., 2011), plants favour sexual reproduction but also maintain the ability to reproduce asexually. These phenomena suggests that sexual reproduction has a slight edge after sacrificing the 50 % cost in general, enough for it to develop and proliferate during the history of evolution. However, a slight increase in its cost could make asexual reproduction more favourable, and more genetic variation (which is restricted by the DNA mutation rate) cannot overcome the disadvantage that only 33 % offspring can reproduce.

If only two random types of many gametes are required in the system, polymorphism of the gametes are almost impossible to evolve. Previous models have predicted a common gamete dimorphism that provides a uniparental inheritance of cytoplasmic genes can offset the costs of mate searching. (Hurst, 1996). Therefore, the gamete dimorphism is also a key factor for the fitness increase in a sexual reproduction.

It has been established that molecular changes can happen before the dimorphism of the gametes (Lipinska et al., 2013). In the case of Zygnematales, the sexual reproduction reversed to a more ancestral-like form, in which genetic materials are exchanged but

there is no gametes dimorphism. DUO1 is also degenerated in this branch. Although argument can be made for "male gametes" no longer exist, the GAMYB proteins does not show such signs. As shown in Chapter 4, DUO1 is essential for the morphogenesis in *Marchantia polymorpha* and *A. thaliana*. Could DUO1 be responsible for the dimorphism (in male, of course) in this entire clade? Further experimental evidence such as a knockout in algae like *Chara* is required before any judgement can be made.

7.3 DUO1 sequence and structural conservation is not always associated with functional importance

There are two regions of DUO1 that show a high conservation in protein sequence, namely the N-terminal MYB domain and the C-terminal TAD motifs. They carry out important functions: the MYB domain for target DNA sequence recognition and binding; the TAD for target gene activation. Certain crucial residues that are required for their functions have been identified previously and in this thesis. Catastrophic effects were observed by disrupting the tryptophans that form the MYB hydrophobic cores or the residues located on the third helices for DNA interaction (Borg, 2011). Similar detrimental effects were also demonstrated on changes of the TAD motifs conserved residues. However, the sequence and structural conservation does not always indicate a functional importance like a self-fulfilling prophecy, nor *vice versa*.

The second helix of the R2 repeat contains a miR159 binding site in most angiosperms, and the specific helical structure was established since bryophytes. Interestingly, a switch at the site to a closely related MYB sequence did not show any negative effect on its function, although the helical structure changed to a more ancestral form. The recruitment of the microRNA regulation also happened much later in the evolution. In this case, the conservation does not conform to a functional importance. On the other hand, changing the diverse regions between the DUO1 and GAMYB protein in *M. polymorpha* failed to produce a MYB domain that binds to the same target sequence as DUO1. Similarly, switching the residues of the MpDUO1 TAD with the AtDUO1 TAD, or swapping the entire TAD region could not raise the target activation ability. These examples suggest that certain regions are functionally important but did not display any conservation observed in this study. This was further demonstrated by deleting the whole AtDUO1

TAD region, which resulted in a molecule that displayed a limited but measureable activation property (Borg, 2011).

7.4 Regulation of the miR159 suggests a potential non-specific DUO1 promoter expression under abiotic stresses

With the current understanding of the microRNA regulation on the target DNA transcriptions, the results are exceptionally puzzling and counterintuitive that the transgene of mAtDUO1 had a lower protein expression *in planta* compared to the AtDUO1.

Assuming the observation was caused by the miR159 regulation, the simplest explanation with available clues is the "bottleneck effect" or "survivorship bias" as discussed in Chapter 5. This hypothesis is based on the highly detrimental effect of the ectopically expressed DUO1 (Palatnik et al., 2007, Brownfield et al., 2009a). The promoter of DUO1 is considered to be male germline-specific (Rotman et al., 2005). In this particular case, the ectopic expression could be the result of the undesired yet frequently observed transgene promoter leakage. Alternatively, the DUO1 promoter could have a weak expression induced by certain stress conditions at an early stage. The ancestral branch of miR159, the miR319 is known to respond to the abiotic stress (Barciszewska-Pacak et al., 2015).

This would also explain the rise of the miR159 binding site in angiosperms. Pre-angiosperm DUO1 homologs do not have the high level of activation abilities observed in their angiosperm counterparts, and they also have a different gene network since genes such as *GCS1* is not under the regulation of DUO1 in *M. polymorpha* (Higo, unpublished). These factors make them potentially less toxic when ectopically expressed. After the fine-tuning of the TAD, this feature became slightly deleterious and the miR159 binding site quickly evolved (from the first angiosperm ca. 150 mya to the first core angiosperm ca. 140 mya).

This hypothesis gives an example of the possibility that expression increase *in vivo* are not always caused by positive regulators, but sometimes by troubles with inhibitors.

7.5 The diversity of the transactivation domains could be an inevitable consequence of evolution as demonstrated by DUO1

The transactivation ability of DUO1 across the eukaryotic kingdoms added another case into such generic transcription regulators. It is counterintuitive that the highly polymorphic linking region would be a fundamental part for the cross-kingdom target activations, even compared to the intrinsic disorder of the mediator complexes (Allen and Taatjes, 2015). The fact that random pieces of molecules can carry out such an important biological function is fascinating yet hard to explain (Erkina and Erkine, 2016). That being said, this phenomenon does seem to agree with the trending idea among biophysicists that life is just the consequence of thermodynamics (Horowitz and England, 2017, Kachman et al., 2017). This "whatever works" or "life finds a way" approach is the reason why we have seen divergent, parallel, and convergent evolution throughout the history of life (Arendt and Reznick, 2008). This goes from as simple as different *Arabidopsis thaliana* lineages adapting to local climate (Stearns and Fenster, 2013), to as complex as cephalopod and human evolved cognitive ability and potentially consciousness separately (Vitti, 2013), or the sexual reproduction in different lineages discussed above. An example exists even in this thesis, where *DUO1* and *GAMYB* converged in recruiting miR159 for regulation separately.

After almost a billion years of evolution, DUO1 has fine-tuned its DNA binding and target activation ability presumably by co-evolving with other parts of the general transcription machinery. However, the core components of DUO1 should have been able to bind to the DNA and activate the targets when first evolved in the Charophyta algae. Like other randomly acquired TADs, this molecule was improved heuristically and any disruption would lead to purification. Now elaborated in fine details, these TADs look unchangeable during the course of evolution, but at their cores are simply the parts that recognise the DNA and start transcription, which will work in any cell using a similar machinery. In the case of this thesis, any eukaryotic cell would satisfy such a condition.

7.6 Future work

DUO1 is a master regulator of plant male germline development. However, the network in which DUO1 functions remains elusive (Borg et al., 2011). Some effort has been put

into the downstream transcription factors DAZ1 and DAZ2 (Borg et al., 2014), yet few other relationships between DUO1 and its targets have been studied. Therefore, such interactions like DUO1-GCS1 or DUO1-GEX2 should be the main focus of any future studies.

That being said, there are some questions still left unanswered about DUO1 itself. The experiments proposed in the discussions of each chapter would resolve some of the issues.

The algal (e.g. *Chara*, *Spirogyra*) DUO1 homologs can be tested using the *in planta* complementation and transient assays, or with the chimeric molecule approach described in Chapter 4. These tests will provide better understanding of the original functions of DUO1 and its link to the male fertility. By detecting the transcripts of the mDUO1 in the WT or DUO1 in the *miR159abc* mutant *Arabidopsis thaliana*, the DUO1 promoter specificity can be confirmed. The Glutathione S-transferase (GST) pull-down assay can be deployed to identify the mediator complex involved in the DUO1 transcriptional regulations, like the study of the AtDreb2a (Aguilar et al., 2014).

Appendix

Table A 1 Sequence of primers used

Primer name	5'-3' sequence
attB1F adaptor	GGGGACAAGTTTGTACAAAAAAGCAGGCT
attB2R adaptor	GGGGACCACTTTGTACAAGAAAGCTGGGT
attB2F adapter	GGGGACAGCTTTCTTGTACAAAGTGG
attB3R adapter	GGGGACAACTTTGTATAATAAAGTTG
attB4F adapter	GGGGACAACTTTGTATAGAAAAGTTG
attB1R adapter	GGGGACTGCTTTTTTGTACAAACTTG
M13F	GTAAAACGACGGCCAG
M13R	CAGGAAACAGCTATGAC
T7	AATACGACTCACTATAGGG
SP6 Promoter R	ATTTAGGTGACACTATAG
BGH Reverse	TAGAAGGCACAGTCGAGG
H2B-att1F	AAAAAGCAGGCTCAATGGCGAAGGCAGATAAGAAACC
H2B-att2R	AGAAAGCTGGGTCCCAGCTCCAGCAGAACTCGTAAAC
GFP-attB2F	TCTTGTACAAAGTGGAAATGAGTAAAGGAGAAGAACTT
GFP-attB3R	TGTATAATAAAGTTGTTATTTGTATAGTTCATCCATGCC
RFP-attB2F	TCTTGTACAAAGTGGCGATGGTGAGCAAGGGCGAGGAGG
RFP-attB3R	TGTATAATAAAGTTGTTTACTTGTACAGCTCGTCCATGC
AtDUO1-fullatt1F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTATGGAAGCGAAGAAGGAAG
AtDUO1-fullatt2Rns	GGGGACCACTTTGTACAAGAAAGCTGGGTAAGGACTTGGGATTGGATCAAC
AtDUO1-fullatt2Rs	GGGGACCACTTTGTACAAGAAAGCTGGCTAAGGACTTGGGATTGGATCAAC
promDUO1-attB4F	TGTATAGAAAAGTTGACGTCCGAAGTTTCCCTCTTGG

promDUO1-attB1R	TTTTGTACAAACTTGCGCTAATCGATCTCTCTCTCG
LlDUO1-attB1F	AAAAAGCAGGCTCGATGGAAGGAGGAGCTGGAG
LlDUO1-attB2R	ACAAGAAAGCTGGGTTCAACCCGGCGGAGGATCCTCC
MpDUO1-attB1F	ACAAAAAAGCAGGCTCTATGAAAACGATTCAAAACGG
MpDUO1-attB2Rstop	ACAAGAAAGCTGGGTACTAGGTTGCCCGACTGAGTTAA
MpDUO1-attB2Rnostop	ACAAGAAAGCTGGGTAGGTTGCCCGACTGAGTTAA
ZmDUO1A-attB1F	ACAAAAAAGCAGGCTCTATGGCTCGACCACCTGG
ZmDUO1A-attB2Rns	ACAAGAAAGCTGGGTAGCTCGTCGCAGGCGG
ZmDUO1A-attB2Rs	ACAAGAAAGCTGGGTACTAGCTCGTCGCAGGCGG
AtDUO1-Chimera-F	CATAACTCCTCTGATGCATCG
PpDUO1-Chimera-R	GCATCAGAGGAGTTATGCAGAGCTCGCAGGATCC
MpDUO1-Chimera-R	GCATCAGAGGAGTTATGCAGTGCGCGCAATATGC
AtDUO1ΔC2-attB2Rs	ACAAGAAAGCTGGGTCCTAGACTTCATCACCAACTTG
AtDUO1ΔC1-attB2Rs	ACAAGAAAGCTGGGTCCTAGTCAAGTCTAGCCAAAAGC
AtDUO1ΔC2-attB2Rns	ACAAGAAAGCTGGGTCGACTTCATCACCAACTTG
AtDUO1ΔC1-attB2Rns	ACAAGAAAGCTGGGTCGTCAAGTCTAGCCAAAAGC
AtDUO1ΔC3-attB2Rns	ACAAGAAAGCTGGGTCCTTTGAAACATGTCTGCATC
AtDUO1ΔC3-attB2Rs	ACAAGAAAGCTGGGTCTCACTTTGAAACATGTCTGCATC
VP16-attB2-R	ACAAGAAAGCTGGGTCCTACCCACCGTACTCGTCAATTCC
VP16ns-attB2R	ACAAGAAAGCTGGGTCCCCACCGTACTCGTCAATTCC
AtDUO1 C-ter KMFK 1	GGATCAACCTGCTTAAACATCTTAGCTGGG
AtDUO1 C-ter KMFK 2	AGGACTTGGGATTGGATCAACCTGCTTA
AtDUO1 C-ter FDSL 1	GGGATTGGATCGAGGGAATCAAACATG
AtDUO1 C-ter FDSL 2	ACAAGAAAGCTGGGTAAGGACTTGGGATTGGATC

AtDUO1 C-ter FLSG 1	CATGTCAGCTGGGAAGCCCGATAAGAAAAC
AtDUO1 C-ter FLSG 1	GGGATTGGATCAACCTGATCAAACATGTCAGC
MpDUO1 C-ter VIDD F	GATAGCGTCATCGACGATTTTGCAGCAG
MpDUO1 C-ter VIDD R	CTGCTGCAAAATCGTCGATGACGCTATC
MpDUO1 C-ter FDQV F	GTGTTTGACCAGGTTGAGCCCCTC
MpDUO1 C-ter FDQV R	GAGGGGCTCAACCTGGTCAAACAC
MpDUO1 C-ter KVFK F	CTTTGCAGCAAAGGTGTTTAAGTCCCTCG
MpDUO1 C-ter KVFK R	CGAGGGACTTAAACACCTTTGCTGCAAAG
AtDUO1 C-terminus F	GATGAAGAAGCTGATG
MpDUO1/AtC-terminus R	CATCAGCTTCTTCATCACTCAGTCGGGCAACCCAC
MpDUO1ΔC1/AtC-terminus R	ACAAGAAAGCTGGGTACCGGGCTACAAGAGTC
MpDUO1ΔC2/AtC-terminus R	CATCAGCTTCTTCATCGTCTTGTCCTCCATCAGTTC
MpDUO1ΔC3/AtC-terminus R	GCATCAGAGGAGTTATGCTGGCTGAAATCGGAATC
MpDUO1ΔC4/AtC-terminus R	CATCAGCTTCTTCATCCGGGCTACAAGAGTC
MpDUO1ΔC4/AtC-terminus F	GGAAGGGTCTTGGATGATC
promHTR10-attB4F	TGTATAGAAAAGTTGTGTTGGCCTATCACGTTGAA
promHTR10-attB1R	TTTTGTACAAACTTGTTCTTCGAGAGAACGATGATG
promDAZ1-attB4F	TGTATAGAAAAGTTGAAGTGGCACAAACCAACCC
promDAZ1-attB1R	TTTTGTACAAACTTGTATTATTGAGTCTCTTACTAGAG
pLeics13-promHTR10:FireLuc-F	TACGGGCCAGATATACGCTGTTGGCCTATCACGTTGAA
pLeics13-promHTR10:FireLuc-R	GACGGAGCTCGAATTTCATTACAATTTGGACTTTCCGCCCT
pLeics12-AtDUO1-F	AGGGAGACCCAAGCTTGGTACCATGGAAGCGAAGAAGGAAG
pLeics12-AtDUO1-R	GACGGAGCTCGAATTTCAAGGACTTGGGATTGGATCAAC
pLeics12-AtDUO1ΔC3-R	GACGGAGCTCGAATTTCACTTTGAAACATGTCTGCATC

pLeics12-AtDUO1ΔC3VP16-R	GACGGAGCTCGAATTTCACCCACCGTACTCGTCAATTCC
PpDUO1A-TA-5'-F	AAAAGGATCCAAATGCAGAATGCTCAGATACATA
PpDUO1A-TA-5'-R	AAAACTCGAGTTGCAAGAAGCCAATCTCATGACA
PpDUO1A-TA-3'-F	AAAAGCGGCCGCGGCTTAATTACCATAACAATGTGC
PpDUO1A-TA-3'-R	AAAATTAATTAAAAATCTACTCCTCTTTTGTCTACC
PpDUO1B-TA-5'-F	AAAAAAGCTTAAATCCATGGCGCCGACAGCTTCC
PpDUO1B-TA-5'-R	AAAACTCGAGTTTGTTTGACCCTGAACCTCCTTG
PpDUO1B-TA-3'-F	AAAAGCGGCCGCTCCAACAACACCAGCAGCAA
PpDUO1B-TA-3'-R	AAAATTAATTAAAAATTCACATGCATACATAGAAGT

Table A 2 Credits for the photographs of the plants

Species	**Source**
Spirogyra	Michael Reese Much FRMS EMS Bethlehem, Pennsylvania, USA. www.photomacrography.net Forum Index -> Photography Through the Microscope
Marchantia polymorpha	NatureSpot. www.naturespot.org.uk/species/marchantia-polymorpha
Selaginella moellendorffii	Lorek, M. 2017. http://www.tropengarten.de/Pflanzen/selaginella-moellendorffii.html
Picea abies	Trees Planet. treesplanet.blogspot.co.uk/2016/05/picea-abies-norway-spruce.html
Amborella trichopoda	Scott Zona. www.flickr.com/photos/scottzona/
Oryza sativa	James Steakley, Wikimedia Commons, Creative Commons Attribution-Share Alike 3.0 Unported license. commons.wikimedia.org/wiki/File:Oryza_sativa_(DITSL).JPG
Arabidopsis thaliana	Dawid Skalec, Wikimedia Commons, Creative Commons Attribution-Share Alike 4.0 International license. commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_2.jpg

Table A 3 Positive selection test results in Poales

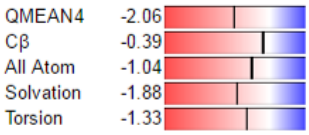
Clade Model (clades): type 0 – background; type 1 – foreground (green box)				
site class	0	1	2	
proportion	0.71990	0.25787	0.02223	
branch type 0	0.04642	1.00000	4.09636	
branch type 1	0.04642	1.00000	4.18866	
Branch-Site Model (individual branches) / Null hypothesis tests				
22-7 PviDUO1B				
site class	0	1	2a	2b
proportion	0.00000	0.00000	0.90022	0.09978
background ω	0.05689	1.00000	0.05689	1.00000
foreground ω	0.05689	1.00000	63.43497	63.43497
24-17 ZmaDUO1C				
site class	0	1	2a	2b
proportion	0.88376	0.09485	0.01931	0.00207
background ω	0.05585	1.00000	0.05585	1.00000
foreground ω	0.05585	1.00000	5.24890	5.24890
26-3 OsaDUO1				
site class	0	1	2a	2b
proportion	0.88005	0.08628	0.03066	0.00301
background ω	0.05681	1.00000	0.05681	1.00000
foreground ω	0.05681	1.00000	40.58886	40.58886
	PST	PST null	2Δ	P Value
22-7	-1886.41708	-1886.51235	0.190538	0.662469
24-17	-1885.64009	-1886.24716	1.21414	0.270514
26-3	-1881.89959	-1886.99430	10.18942	0.001412

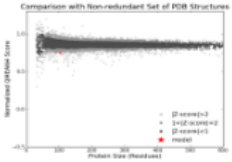
Table A 4 Accession numbers and sources of DUO1 homolog sequences

Species	Gene name	Source	Gene Name (Based on the Source)
Amaranthus hypochondriacus	AhyDUO1A	Phytozome	AHYPO_006386
	AhyDUO1B	Phytozome	AHYPO_014958
Amborella trichopoda	AtrDUO1	NCBI	LOC18442583
Ananas comosus	AcoDUO1	Phytozome	Aco012841
Aquilegia caerulea	AcaDUO1A	Phytozome	Aquca_002_00029.1
	AcaDUO1B	Phytozome	Aquca_081_00025.1
Arabidopsis halleri	AhaDUO1	Phytozome	Araha.58829s0002.1
Arabidopsis lyrata	AlyDUO1	Phytozome	486519
Arabidopsis thaliana	AthDUO1	TAIR	AT3G60460.1
Boechera stricta	BstDUO1	Phytozome	Bostr.13158s0042
Brachypodium distachyon	BdiDUO1	Phytozome	Bradi5g17600.1
Brassica rapa	BraDUO1A	Phytozome	Brara.G01884.1
	BraDUO1B	Phytozome	Brara.I04172.1
Capsella grandiflora	CgrDUO1	Phytozome	Cagra.3527s0008.1
Capsella rubella	CruDUO1	Phytozome	Carubv10019483m
Carica papaya	CpaDUO1	Phytozome	evm.model.supercontig_37.96
Citrus clementina	CclDUO1	Phytozome	Ciclev10021345m
Citrus sinensis	CsiDUO1	Phytozome	orange1.1g038795m
Cucumis sativus	CsaDUO1	Phytozome	Cucsa.094180.1
Daucus carota	DcaDUO1	Phytozome	DCAR_029776
Eucalyptus grandis	EgrDUO1	Phytozome	Eucgr.E01581.1
Fragaria vesca	FveDUO1A	Phytozome	mrna04800.1
	FveDUO1B	Phytozome	mrna12261.1
Glycine max	GmaDUO1A	Phytozome	Glyma03g26830.2
	GmaDUO1B	Phytozome	Glyma07g14480.1
	GmaDUO1C	Phytozome	Glyma09g00371.1
	GmaDUO1D	Phytozome	Glyma12g37026.1
Gossypium raimondii	GraDUO1	Phytozome	Gorai.004G165400.1
Kalanchoe fedtschenkoi	KfeDUO1A	Phytozome	Kaladp0068s0290.1
	KfeDUO1B	Phytozome	Kaladp0079s0022.1
	KfeDUO1C	Phytozome	Kaladp0103s0001.1
Kalanchoe laxiflora	KlaDUO1A	Phytozome	Kalax.0004s0050.1
	KlaDUO1B	Phytozome	Kalax.0021s0115.1
	KlaDUO1C	Phytozome	Kalax.0415s0009.1
	KlaDUO1D	Phytozome	Kalax.0473s0007.1
	KlaDUO1E	Phytozome	Kalax.0755s0018.1
Lilium longiflorum	LloDUO1	cDNA	N/A
Linum usitatissimum	LusDUO1	Phytozome	Lus10009780
Malus domestica	MdoDUO1	Phytozome	MDP0000237596
Manihot esculenta	MesDUO1A	Phytozome	cassava4.1_021857m
	MesDUO1B	Phytozome	cassava4.1_022856m
Medicago truncatula	MtrDUO1	Phytozome	Medtr8g006470.1
Mimulus guttatus	MguDUO1A	Phytozome	mgv1a012916m
	MguDUO1B	Phytozome	mgv1a022833m

Marchantia polymorpha	MpoDUO1	Phytozome	Mapoly0019s0071.1
Musa acuminata	MacDUO1	Phytozome	GSMUA_Achr1G01660_001
Nicotiana tabacum	NtaDUO1	NCBI	LOC107810051
Oropetium thomaeum	OthDUO1	Phytozome	Oropetium_20150105_12296A
Oryza sativa	OsaDUO1	Phytozome	LOC_Os04g46384.1
Panicum hallii	PhaDUO1A	Phytozome	Pahal.G01829.1
	PhaDUO1B	Phytozome	Pahal.F00780.1
Panicum virgatum	PviDUO1A	Phytozome	Pavirv00018278m
	PviDUO1B	Phytozome	Pavirv00041069m
	PviDUO1C	Phytozome	Pavirv00015512m
	PviDUO1D	Phytozome	Pavirv00020433m
Phaseolus vulgaris	PvuDUO1A	Phytozome	Phvul.010G053200.1
	PvuDUO1B	Phytozome	Phvul.011G215300.1
Physcomitrella patens	PpaDUO1A	Ensembl Plant	PP1S16_281V6
	PpaDUO1B	Ensembl Plant	PP1S114_136V6
Picea abies	PabDUO1	Congenie.org	MA_130648g0010
Populus trichocarpa	PtrDUO1A	Phytozome	Potri.002G140900.1
	PtrDUO1B	Phytozome	Potri.014G054700.1
Prunus persica	PpeDUO1A	Phytozome	ppa016603m
	PpeDUO1B	Phytozome	ppa020326m
Ricinus communis	RcoDUO1	Phytozome	30174.m008784
Salix purpurea	SpuDUO1A	Phytozome	SapurV1A.0033s0200
	SpuDUO1B	Phytozome	SapurV1A.0490s0170
Selaginella moellendorffii	SmoDUO1A	Ensembl Plant	SELMODRAFT_39442
	SmoDUO1B	Ensembl Plant	SELMODRAFT_80215
Setaria italica	SitDUO1A	Phytozome	Si012032m
	SitDUO1B	Phytozome	Si015815m
Solanum lycopersicum	SlyDUO1A	Phytozome	Solyc01g090530.1.1
	SlyDUO1B	Phytozome	Solyc10g019260.1.1
Solanum tuberosum	StuDUO1	Phytozome	PGSC0003DMG400025720
Sorghum bicolor	SbiDUO1A	Phytozome	Sb06g024510.1
	SbiDUO1B	Phytozome	Sb07g027160.1
Sphagnum fallax	SfaDUO1	Phytozome	Sphfalx0015s0253.1
Spirodela polyrhiza	SpoDUO1	Phytozome	Spipo7G0012500
Thellungiella halophila	ThaDUO1	Phytozome	Thhalv10006475m
Theobroma cacao	TcaDUO1A	Phytozome	Thecc1EG014872t1
	TcaDUO1B	Phytozome	Thecc1EG005653t1
Trifolium pratense	TprDUO1	Phytozome	Tp57577_TGAC_v2_gene24946
Triticum aestivum	TaeDUO1A	Phytozome	Traes_2BL_855A1170C.1
	TaeDUO1B	Phytozome	Traes_2AL_0A21FB42C.1
	TaeDUO1C	Phytozome	Traes_2DL_912473A86.1
Vitis vinifera	VviDUO1	Phytozome	GSVIVT01018234001
Zea mays	ZmaDUO1A	Phytozome	GRMZM2G105137_T01
	ZmaDUO1B	Phytozome	GRMZM2G046443_T01
	ZmaDUO1C	Phytozome	GRMZM2G311059_T01
Zostera marina	ZmaDUO1	Phytozome	Zosma38g00960

Model #01	File	Built with	Oligo-State	Ligands	GMQE	QMEAN4
	PDB	ProMod Version 3.70.	MONOMER	None	0.80	-2.06





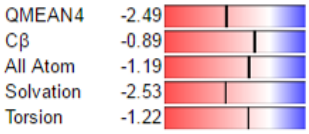
Template	Seq Identity	Oligo-state	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
1h8a.1.C	49.02	hetero-oligomer	BLAST	X-ray	2.23Å	0.46		0.97	MYB TRANSFORMING PROTEIN

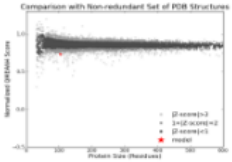
The template contained no ligands.

```
Target    IKKGPWKAEEDEVLINHVKRYGPRDWSSIRSKGLLQRTGKSCRLRWVNKLRPNLKNGCKFSADEERTVIELQSEFGNKWA
1h8a.1.C  LNKGPWTKEEDQRVIEHVQKYGPKRWSDI-AKHLKGRIGKQCRERWHNHLNPEVKK-TSWTEEEDRIIYQAHKRLGNRWA

Target    RIATYLPGRTDNDVKNFWSSRQKRL
1h8a.1.C  EIAKLLPGRTDNAVKNHWNSTMRR-
```

Model #02	File	Built with	Oligo-State	Ligands	GMQE	QMEAN4
	PDB	ProMod Version 3.70.	MONOMER	None	0.78	-2.49





Template	Seq Identity	Oligo-state	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
1h8a.1.C	48.04	hetero-oligomer	HHblits	X-ray	2.23Å	0.45		0.97	MYB TRANSFORMING PROTEIN

The template contained no ligands.

```
Target    IKKGPWKAEEDEVLINHVKRYGPRDWSSIRSKGLLQRTGKSCRLRWVNKLRPNLKNGCKFSADEERTVIELQSEFGNKWA
1h8a.1.C  LNKGPWTKEEDQRVIEHVQKYGPKRWSDIAK-HLKGRIGKQCRERWHNHLNPEVKKT-SWTEEEDRIIYQAHKRLGNRWA

Target    RIATYLPGRTDNDVKNFWSSRQKRL
1h8a.1.C  EIAKLLPGRTDNAVKNHWNSTMRR-
```

Figure A 1 Parameters for the two protein models of the DUO1 MYB domain.
The models were generated using SWISS-MODEL based on the template 1h8a.1.C. A single amino acid alignment difference at the B region (GCK) resulted in two different models.

Figure A 2 Spirogyra GAMYB homologs that show no sign of degeneration.

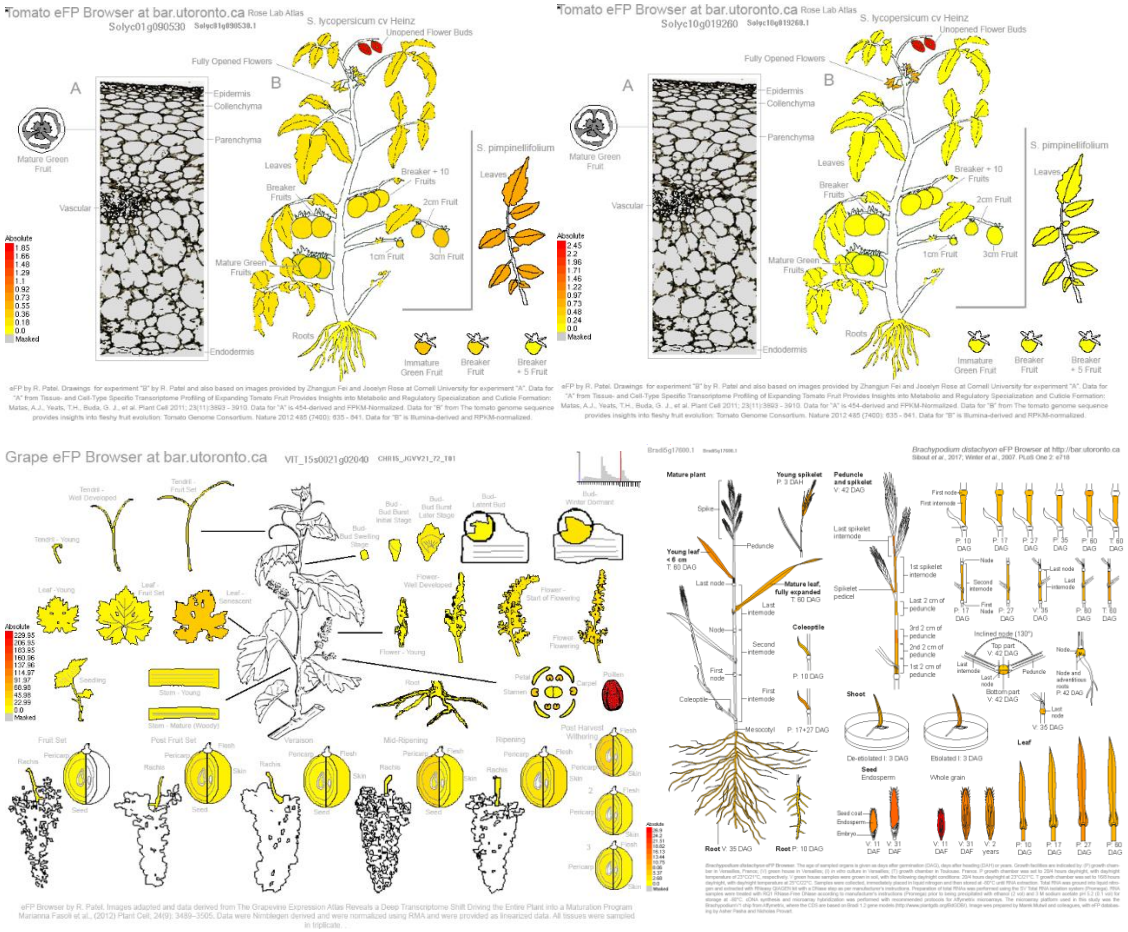


Figure A 3 DUO1 homolog expression patterns from BAR website.
These include DUO1 homologs from Solanum lycopersicum, Vitis vinifera, and Oryza sativa. They all show a male tissue specific expression.

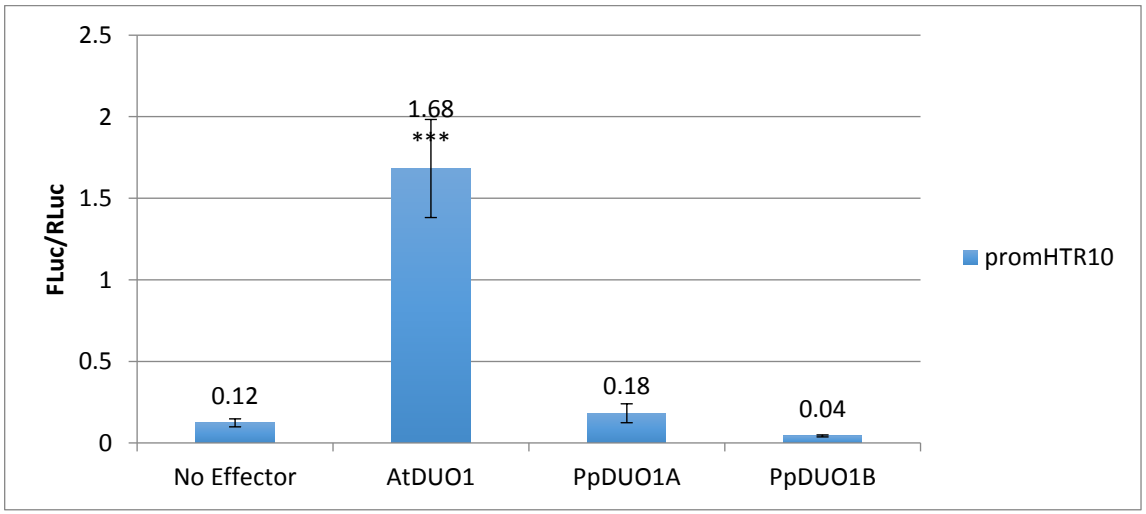


Figure A 4 Independent repeat of *Physcomitrella patans* DUO1 homologs.
The homologs were tested in the transient assays using the AtHTR10 promoter. The results were similar to the ones shown in Chapter 4. Experiments conducted by Mikhaela Neequaye.

	Forward	Reverse	Size in WT	Size in PpDUO1A_KO
WT	AL006 acatgcatcgaaaaagaacg	AL007 tttgttctggctgcatcatt	1127 bp	-
G418 casette	CR7 CGAGCTCGAATTCCCATGGA	CT8 GCAAGGTGAGATGACAGGAGAT	-	1005 bp
5´integration	AL006 acatgcatcgaaaaagaacg	oAT44 TACGGCGAGTTCTGTTAGGTC	-	1100 bp
3´integration	AL028 AGGGTTTCGCTCATGTGTTG	AL008 caatagataccgttatacaaact	-	1145 bp
Full	AL006 acatgcatcgaaaaagaacg	AL008 caatagataccgttatacaaact	4343 bp	3753 bp

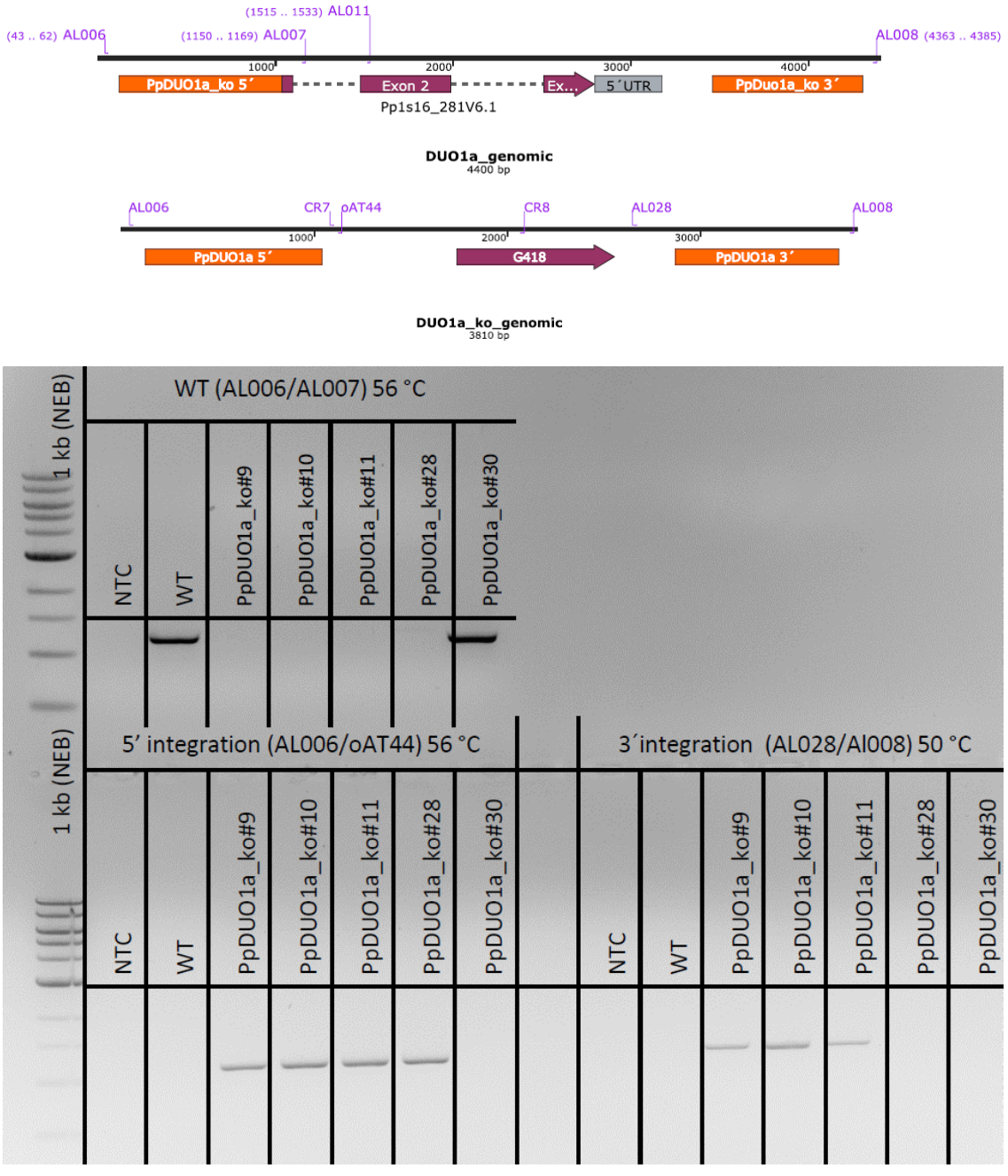


Figure A 5 Genotyping of knockout in moss.
The genotyping was designed and carried out by the collaborator Dr Ann-Cathrin Lindner from Jörg Becker's group in IGC, Portugal.

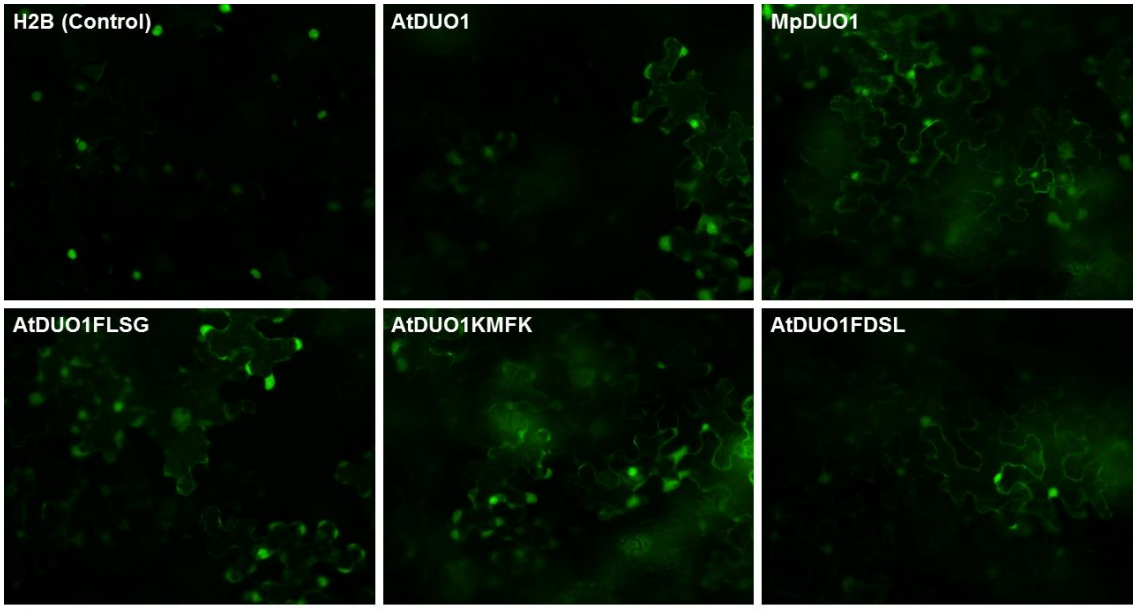


Figure A 6 GFP signals of the C-terminus variants.
No obvious differences in the GFP expression level was observed for these variants. Similar results were found in the MpDUO1 variants.

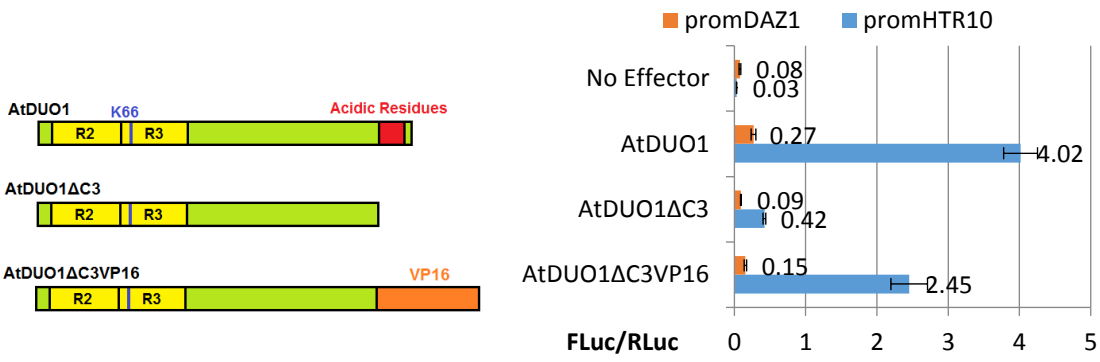


Figure A 7 The activation ability of VP16 for the AtDAZ1 promoter.
As was shown in the case of HTR10 promoter in the transient assay, VP16 was able to rescue the activation defect of the AtDUO1ΔC3 variant.

Bibliography

ACHARD, P., HERR, A., BAULCOMBE, D. C. & HARBERD, N. P. 2004. Modulation of floral development by a gibberellin-regulated microRNA. *Development,* 131**,** 3357-65.

ADAI, A., JOHNSON, C., MLOTSHWA, S., ARCHER-EVANS, S., MANOCHA, V., VANCE, V. & SUNDARESAN, V. 2005. Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res,* 15**,** 78-91.

ADKINS, N. L., HAGERMAN, T. A. & GEORGEL, P. 2006. GAGA protein: a multi-faceted transcription factor. *Biochem Cell Biol,* 84**,** 559-67.

AGUILAR, X., BLOMBERG, J., BRANNSTROM, K., OLOFSSON, A., SCHLEUCHER, J. & BJORKLUND, S. 2014. Interaction studies of the human and Arabidopsis thaliana Med25-ACID proteins with the herpes simplex virus VP16- and plant-specific Dreb2a transcription factors. *PLoS One,* 9**,** e98575.

AHMED, S., COCK, J. M., PESSIA, E., LUTHRINGER, R., CORMIER, A., ROBUCHON, M., STERCK, L., PETERS, A. F., DITTAMI, S. M., CORRE, E., VALERO, M., AURY, J. M., ROZE, D., VAN DE PEER, Y., BOTHWELL, J., MARAIS, G. A. & COELHO, S. M. 2014. A haploid system of sex determination in the brown alga Ectocarpus sp. *Curr Biol,* 24**,** 1945-57.

ALBERT, V. A., BARBAZUK, W. B., DEPAMPHILIS, C. W., DER, J. P., LEEBENS-MACK, J., MA, H., PALMER, J. D., ROUNSLEY, S., SANKOFF, D., SCHUSTER, S. C., SOLTIS, D. E., SOLTIS, P. S., WESSLER, S. R., WING, R. A., ALBERT, V. A., AMMIRAJU, J. S. S., BARBAZUK, W. B., CHAMALA, S., CHANDERBALI, A. S., DEPAMPHILIS, C. W., DER, J. P., DETERMANN, R., LEEBENS-MACK, J., MA, H., RALPH, P., ROUNSLEY, S., SCHUSTER, S. C., SOLTIS, D. E., SOLTIS, P. S., TALAG, J., TOMSHO, L., WALTS, B., WANKE, S., WING, R. A., ALBERT, V. A., BARBAZUK, W. B., CHAMALA, S., CHANDERBALI, A. S., CHANG, T. H., DETERMANN, R., LAN, T. Y., SOLTIS, D. E., SOLTIS, P. S., ARIKIT, S., AXTELL, M. J., AYYAMPALAYAM, S., BARBAZUK, W. B., BURNETTE, J. M., CHAMALA, S., DE PAOLI, E., DEPAMPHILIS, C. W., DER, J. P., ESTILL, J. C., FARRELL, N. P., HARKESS, A., JIAO, Y. N., LEEBENS-MACK, J., LIU, K., MEI, W. B., MEYERS, B. C., SHAHID, S., WAFULA, E., WALTS, B., WESSLER, S. R., ZHAI, J. X., ZHANG, X. Y., ALBERT, V. A., CARRETERO-PAULET, L., DEPAMPHILIS, C. W., DER, J. P., JIAO, Y. N., LEEBENS-MACK, J., LYONS, E., SANKOFF, D., TANG, H. B., WAFULA, E., ZHENG, C. F., ALBERT, V. A., ALTMAN, N. S., BARBAZUK, W. B., CARRETERO-PAULET, L., DEPAMPHILIS, C. W., DER, J. P., ESTILL, J. C., JIAO, Y. N., LEEBENS-MACK, J., LIU, K., MEI, W. B., WAFULA, E., ALTMAN, N. S., ARIKIT, S., AXTELL, M. J., CHAMALA, S., CHANDERBALI, A. S., CHEN, F., CHEN, J. Q., CHIANG, V., DE PAOLI, E., DEPAMPHILIS, C. W., DER, J. P., et al. 2013. The Amborella Genome and the Evolution of Flowering Plants. *Science,* 342**,** 1467-+.

ALLEN, B. L. & TAATJES, D. J. 2015. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol,* 16**,** 155-66.

ALLEN, R. S., LI, J., ALONSO-PERAL, M. M., WHITE, R. G., GUBLER, F. & MILLAR, A. A. 2010. MicroR159 regulation of most conserved targets in Arabidopsis has negligible phenotypic effects. *Silence,* 1**,** 18.

ALLEN, R. S., LI, J., STAHLE, M. I., DUBROUE, A., GUBLER, F. & MILLAR, A. A. 2007. Genetic analysis reveals functional redundancy and the major target genes of the Arabidopsis miR159 family. *Proc Natl Acad Sci U S A,* 104**,** 16371-6.

ALONSO-PERAL, M. M., LI, J., LI, Y., ALLEN, R. S., SCHNIPPENKOETTER, W., OHMS, S., WHITE, R. G. & MILLAR, A. A. 2010. The microRNA159-regulated GAMYB-like genes inhibit growth and promote programmed cell death in Arabidopsis. *Plant Physiol,* 154**,** 757-71.

ALONSO-PERAL, M. M., SUN, C. & MILLAR, A. A. 2012. MicroRNA159 Can Act as a Switch or Tuning MicroRNA Independently of Its Abundance in Arabidopsis. *Plos One,* 7.

ALVES-FERREIRA, M., WELLMER, F., BANHARA, A., KUMAR, V., RIECHMANN, J. L. & MEYEROWITZ, E. M. 2007. Global expression profiling applied to the analysis of Arabidopsis stamen development. *Plant Physiol,* 145**,** 747-62.

ALVES, L., JR., NIEMEIER, S., HAUENSCHILD, A., REHMSMEIER, M. & MERKLE, T. 2009. Comprehensive prediction of novel microRNA targets in Arabidopsis thaliana. *Nucleic Acids Res,* 37**,** 4010-21.

AN, R., LIU, X., WANG, R., WU, H., LIANG, S., SHAO, J., QI, Y., AN, L. & YU, F. 2014. The over-expression of two transcription factors, ABS5/bHLH30 and ABS7/MYB101, leads to upwardly curly leaves. *PLoS One,* 9**,** e107637.

ANDERSON, S. N., JOHNSON, C. S., JONES, D. S., CONRAD, L. J., GOU, X., RUSSELL, S. D. & SUNDARESAN, V. 2013. Transcriptomes of isolated Oryza sativa gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J,* 76**,** 729-41.

ANDREJKA, L., WEN, H., ASHTON, J., GRANT, M., IORI, K., WANG, A., MANAK, J. R. & LIPSICK, J. S. 2011. Animal-specific C-terminal domain links myeloblastosis oncoprotein (Myb) to an ancient repressor complex. *Proc Natl Acad Sci U S A,* 108**,** 17438-43.

AOYAMA, T., HIWATASHI, Y., SHIGYO, M., KOFUJI, R., KUBO, M., ITO, M. & HASEBE, M. 2012. AP2-type transcription factors determine stem cell identity in the moss Physcomitrella patens. *Development,* 139**,** 3120-9.

ARAKI, S., ITO, M., SOYANO, T., NISHIHAMA, R. & MACHIDA, Y. 2004. Mitotic cyclins stimulate the activity of c-Myb-like factors for transactivation of G2/M phase-specific genes in tobacco. *J Biol Chem,* 279**,** 32979-88.

ARENDT, J. & REZNICK, D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution,* 23**,** 26-32.

ARNOLD, K., BORDOLI, L., KOPP, J. & SCHWEDE, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* 22**,** 195-201.

ASHMAN, T. L., BACHTROG, D., BLACKMON, H., GOLDHERG, E. E., HAHN, M. W., KIRKPATRICK, M., KITANO, J., MANK, J. E., MAYROSE, I., MING, R., OTTO, S. P., PEICHEL, C. L., PENNELL, M. W., PERRIN, N., ROSS, L., VALENZUELA, N., VAMOSI, J. C. & CONSORTIUM, T. S. 2014. Tree of Sex: A database of sexual systems. *Scientific Data,* 1.

ASHTON, N. W., GRIMSLEY, N. H. & COVE, D. J. 1979. Analysis of gametophytic development in the moss, Physcomitrella patens, using auxin and cytokinin resistant mutants. *Planta,* 144**,** 427-35.

AUKERMAN, M. J. & SAKAI, H. 2003. Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell,* 15**,** 2730-41.

AVILA, J., NIETO, C., CANAS, L., BENITO, M. J. & PAZ-ARES, J. 1993. Petunia hybrida genes related to the maize regulatory C1 gene and to animal myb proto-oncogenes. *Plant J,* 3**,** 553-62.

AXTELL, M. J. & BOWMAN, J. L. 2008. Evolution of plant microRNAs and their targets. *Trends Plant Sci,* 13**,** 343-9.

AXTELL, M. J., SNYDER, J. A. & BARTEL, D. P. 2007. Common functions for diverse small RNAs of land plants. *Plant Cell,* 19**,** 1750-69.

AYA, K., UEGUCHI-TANAKA, M., KONDO, M., HAMADA, K., YANO, K., NISHIMURA, M. & MATSUOKA, M. 2009. Gibberellin modulates anther development in rice via the transcriptional regulation of GAMYB. *Plant Cell,* 21**,** 1453-72.

AYORA, S., PIRUAT, J. I., LUNA, R., REISS, B., RUSSO, V. E. A., AGUILERA, A. & ALONSO, J. C. 2002. Characterization of two highly similar Rad51 homologs of Physcomitrella patens. *Journal of Molecular Biology,* 316**,** 35-49.

BABU, M. M., LUSCOMBE, N. M., ARAVIND, L., GERSTEIN, M. & TEICHMANN, S. A. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol,* 14**,** 283-91.

BALLESTEROS, M. L., BOLLE, C., LOIS, L. M., MOORE, J. M., VIELLE-CALZADA, J. P., GROSSNIKLAUS, U. & CHUA, N. H. 2001. LAF1, a MYB transcription activator for phytochrome A signaling. *Genes Dev,* 15**,** 2613-25.

BARCISZEWSKA-PACAK, M., MILANOWSKA, K., KNOP, K., BIELEWICZ, D., NUC, P., PLEWKA, P., PACAK, A. M., VAZQUEZ, F., KARLOWSKI, W., JARMOLOWSKI, A. & SZWEYKOWSKA-KULINSKA, Z. 2015. Arabidopsis microRNA expression regulation in a wide range of abiotic stress responses. *Front Plant Sci,* 6**,** 410.

BARKER, M. S., VOGEL, H. & SCHRANZ, M. E. 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome Biol Evol,* 1**,** 391-9.

BARTEL, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-97.

BATEMAN, A. J. 1948. Intra-sexual selection in Drosophila. *Heredity (Edinb),* 2**,** 349-68.

BAUMANN, P. & WEST, S. C. 1998. Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. *Trends Biochem Sci,* 23**,** 247-51.

BEALL, E. L., MANAK, J. R., ZHOU, S., BELL, M., LIPSICK, J. S. & BOTCHAN, M. R. 2002. Role for a Drosophila Myb-containing protein complex in site-specific DNA replication. *Nature,* 420**,** 833-7.

BERGER, F. & TWELL, D. 2011. Germline specification and function in plants. *Annu Rev Plant Biol,* 62**,** 461-84.

BERMAN, H. M. 2008. The Protein Data Bank: a historical perspective. *Acta Crystallogr A,* 64**,** 88-95.

BERMAN, H. M., KLEYWEGT, G. J., NAKAMURA, H. & MARKLEY, J. L. 2014. The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des*.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res,* 28**,** 235-42.

BHATT, A. M., CANALES, C. & DICKINSON, H. G. 2001. Plant meiosis: the means to 1N. *Trends in Plant Science,* 6**,** 114-121.

BHATTACHARYA, D. & MEDLIN, L. 1998. Algal phylogeny and the origin of land plants. *Plant Physiology,* 116**,** 9-15.

BLACK, D. L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem,* 72**,** 291-336.

BLANC, G., HOKAMP, K. & WOLFE, K. H. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res,* 13**,** 137-44.

BLOMME, T., VANDEPOELE, K., DE BODT, S., SIMILLION, C., MAERE, S. & VAN DE PEER, Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol,* 7**,** R43.

BONNET, E., HE, Y., BILLIAU, K. & VAN DE PEER, Y. 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics,* 26**,** 1566-8.

BOOTH, W., JOHNSON, D. H., MOORE, S., SCHAL, C. & VARGO, E. L. 2011. Evidence for viable, non-clonal but fatherless Boa constrictors. *Biology Letters,* 7**,** 253-256.

BORG, M. 2011. *Regulatory mechanisms controlling male germline development in Arabidopsis thaliana [electronic resource]* [Online]. University of Leicester. Available: http://hdl.handle.net/2381/10081.

BORG, M. & BERGER, F. 2015. Chromatin remodelling during male gametophyte development. *Plant J,* 83**,** 177-88.

BORG, M., BROWNFIELD, L., KHATAB, H., SIDOROVA, A., LINGAYA, M. & TWELL, D. 2011. The R2R3 MYB transcription factor DUO1 activates a male germline-specific regulon essential for sperm cell differentiation in Arabidopsis. *Plant Cell,* 23**,** 534-49.

BORG, M., BROWNFIELD, L. & TWELL, D. 2009. Male gametophyte development: a molecular perspective. *Journal of Experimental Botany,* 60**,** 1465-1478.

BORG, M., RUTLEY, N., KAGALE, S., HAMAMURA, Y., GHERGHINOIU, M., KUMAR, S., SARI, U., ESPARZA-FRANCO, M. A., SAKAMOTO, W., ROZWADOWSKI, K., HIGASHIYAMA, T. & TWELL, D. 2014. An EAR-Dependent Regulatory Module Promotes Male Germ Cell Division and Sperm Fertility in Arabidopsis. *Plant Cell,* 26**,** 2098-2113.

BORTESI, L. & FISCHER, R. 2015. The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol Adv,* 33**,** 41-52.

BOSTAN, H. & CHIUSANO, M. L. 2015. NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. *BMC Plant Biol,* 15**,** 48.

BOWERS, J. E., CHAPMAN, B. A., RONG, J. & PATERSON, A. H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature,* 422**,** 433-8.

BOWMAN, J. L. 2015. A Brief History of Marchantia from Greece to Genomics. *Plant Cell Physiol*.

BOWMAN, J. L., KOHCHI, T., YAMATO, K. T., JENKINS, J., SHU, S. Q., ISHIZAKI, K., YAMAOKA, S., NISHIHAMA, R., NAKAMURA, Y., BERGER, F., ADAM, C., AKI, S. S., ALTHOFF, F., ARAKI, T., ARTEAGA-VAZQUEZ, M. A., BALASUBRMANIAN, S., BARRY, K., BAUER, D., BOEHM, C. R., BRIGINSHAW, L., CABALLERO-PEREZ, J., CATARINO, B., CHEN, F., CHIYODA, S., CHOVATIA, M., DAVIES, K. M., DELMANS, M., DEMURA, T., DIERSCHKE, T., DOLAN, L., DORANTES-ACOSTA, A. E., EKLUND, D. M., FLORENT, S. N., FLORES-SANDOVAL, E., FUJIYAMA, A., FUKUZAWA, H., GALIK, B., GRIMANELLI, D., GRIMWOOD, J., GROSSNIKLAUS, U., HAMADA, T., HASELOFF, J., HETHERINGTON, A. J., HIGO, A., HIRAKAWA, Y., HUNDLEY, H. N., IKEDA, Y., INOUE, K., INOUE, S. I., ISHIDA, S., JIA, Q. D., KAKITA, M., KANAZAWA, T., KAWAI, Y., KAWASHIMA, T., KENNEDY, M., KINOSE, K., KINOSHITA, T., KOHARA, Y., KOIDE, E., KOMATSU, K., KOPISCHKE, S., KUBO, M., KYOZUKA, J., LAGERCRANTZ, U., LIN, S. S., LINDQUIST, E., LIPZEN, A. M., LU, C. W., DE LUNA, E., MARTIENSSEN, R. A., MINAMINO, N., MIZUTANI, M., MIZUTANI, M., MOCHIZUKI, N., MONTE, I., MOSHER, R., NAGASAKI, H., NAKAGAMI, H., NARAMOTO, S., NISHITANI, K., OHTANI, M., OKAMOTO, T., OKUMURA, M., PHILLIPS, J., POLLAK, B., REINDERS, A., ROVEKAMP, M., SANO, R., SAWA, S., SCHMID, M. W., SHIRAKAWA, M., SOLANO, R., SPUNDE, A., SUETSUGU, N., SUGANO, S., SUGIYAMA, A., SUN, R., SUZUKI, Y., TAKENAKA, M., et al. 2017. Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell,* 171**,** 287-+.

BOYES, D. C., ZAYED, A. M., ASCENZI, R., MCCASKILL, A. J., HOFFMAN, N. E., DAVIS, K. R. & GORLACH, J. 2001. Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell,* 13**,** 1499-510.

BRADY, A. & SALZBERG, S. 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat Methods,* 8**,** 367.

BRAUN, E. L. & GROTEWOLD, E. 1999. Newly discovered plant c-myb-like genes rewrite the evolution of the plant myb gene family. *Plant Physiol,* 121**,** 21-4.

BREMER, B., BREMER, K., CHASE, M. W., FAY, M. F., REVEAL, J. L., SOLTIS, D. E., SOLTIS, P. S., STEVENS, P. F., ANDERBERG, A. A., MOORE, M. J., OLMSTEAD, R. G., RUDALL, P. J., SYTSMA, K. J., TANK, D. C., WURDACK, K., XIANG, J. Q. Y., ZMARZTY, S. & GRP, A. P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society,* 161**,** 105-121.

BREMER, B., BREMER, K., CHASE, M. W., REVEAL, J. L., SOLTIS, D. E., SOLTIS, P. S., STEVENS, P. F., ANDERBERG, A. A., FAY, M. F., GOLDBLATT, P., JUDD, W. S., KALLERSJO, M., KAREHED, J., KRON, K. A., LUNDBERG, J., NICKRENT, D. L., OLMSTEAD, R. G., OXELMAN, B., PIRES,

J. C., RODMAN, J. E., RUDALL, P. J., SAVOLAINEN, V., SYTSMA, K. J., VAN DER BANK, M., WURDACK, K., XIANG, J. Q. Y., ZMARZTY, S. & GRP, A. P. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society,* 141**,** 399-436.

BRENNAN, R. G. & MATTHEWS, B. W. 1989. The helix-turn-helix DNA binding motif. *J Biol Chem,* 264**,** 1903-6.

BRENNECKE, J., HIPFNER, D. R., STARK, A., RUSSELL, R. B. & COHEN, S. M. 2003. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell,* 113**,** 25-36.

BRIVANLOU, A. H. & DARNELL, J. E., JR. 2002. Signal transduction and the control of gene expression. *Science,* 295**,** 813-8.

BROWN, G. R., LALAND, K. N. & MULDER, M. B. 2009. Bateman's principles and human sex roles. *Trends Ecol Evol,* 24**,** 297-304.

BROWNFIELD, L., HAFIDH, S., BORG, M., SIDOROVA, A., MORI, T. & TWELL, D. 2009a. A plant germline-specific integrator of sperm specification and cell cycle progression. *PLoS Genet,* 5**,** e1000430.

BROWNFIELD, L., HAFIDH, S., DURBARRY, A., KHATAB, H., SIDOROVA, A., DOERNER, P. & TWELL, D. 2009b. Arabidopsis DUO POLLEN3 is a key regulator of male germline development and embryogenesis. *Plant Cell,* 21**,** 1940-56.

BURATOWSKI, S., HAHN, S., GUARENTE, L. & SHARP, P. A. 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell,* 56**,** 549-61.

BURLEIGH, J. G., BANSAL, M. S., EULENSTEIN, O., HARTMANN, S., WEHE, A. & VISION, T. J. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. *Syst Biol,* 60**,** 117-25.

BURLEIGH, J. G., HILU, K. W. & SOLTIS, D. E. 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. *BMC Evol Biol,* 9**,** 61.

BURT, A. 2000. Perspective: Sex, recombination, and the efficacy of selection - Was Weismann right? *Evolution,* 54**,** 337-351.

BUXDORF, K., HENDELMAN, A., STAV, R., LAPIDOT, M., ORI, N. & ARAZI, T. 2010. Identification and characterization of a novel miR159 target not related to MYB in tomato. *Planta,* 232**,** 1009-22.

BYNG, J. W., CHASE, M. W., CHRISTENHUSZ, M. J. M., FAY, M. F., JUDD, W. S., MABBERLEY, D. J., SENNIKOV, A. N., SOLTIS, D. E., SOLTIS, P. S., STEVENS, P. F., BRIGGS, B., BROCKINGTON, S., CHAUTEMS, A., CLARK, J. C., CONRAN, J., HASTON, E., MOLLER, M., MOORE, M., OLMSTEAD, R., PERRET, M., SKOG, L., SMITH, J., TANK, D., VORONTSOVA, M., WEBER, A. & GRP, A. P. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society,* 181**,** 1-20.

CAPECCHI, M. R. 1989. Altering the genome by homologous recombination. *Science,* 244**,** 1288-92.

CEDRONI, M. L., CRONN, R. C., ADAMS, K. L., WILKINS, T. A. & WENDEL, J. F. 2003. Evolution and expression of MYB genes in diploid and polyploid cotton. *Plant Mol Biol,* 51**,** 313-25.

CHANG, Y. & GRAHAM, S. W. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am J Bot,* 98**,** 839-49.

CHARLOT, F., CHELYSHEVA, L., KAMISUGI, Y., VRIELYNCK, N., GUYON, A., EPERT, A., LE GUIN, S., SCHAEFER, D. G., CUMING, A. C., GRELON, M. & NOGUE, F. 2014. RAD51B plays an essential role during somatic and meiotic recombination in Physcomitrella. *Nucleic Acids Res,* 42**,** 11965-78.

CHEN, X. 2004. A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science,* 303**,** 2022-5.

CHENG, H., QIN, L., LEE, S., FU, X., RICHARDS, D. E., CAO, D., LUO, D., HARBERD, N. P. & PENG, J. 2004. Gibberellin regulates Arabidopsis floral development via suppression of DELLA protein function. *Development,* 131**,** 1055-64.

CHETTOOR, A. M., GIVAN, S. A., COLE, R. A., COKER, C. T., UNGER-WALLACE, E., VEJLUPKOVA, Z., VOLLBRECHT, E., FOWLER, J. E. & EVANS, M. M. 2014. Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome Biol,* 15**,** 414.

CHIYODA, S., ISHIZAKI, K., KATAOKA, H., YAMATO, K. T. & KOHCHI, T. 2008. Direct transformation of the liverwort Marchantia polymorpha L. by particle bombardment using immature thalli developing from spores. *Plant Cell Reports,* 27**,** 1467-1473.

CHIYODA, S., LINLEY, P. J., YAMATO, K. T., FUKUZAWA, H., YOKOTA, A. & KOHCHI, T. 2006. Simple and efficient plastid transformation system for the liverwort Marchantia polymorpha L. suspension-culture cells. *Transgenic Research,* 16**,** 41-49.

CIVAN, P., FOSTER, P. G., EMBLEY, M. T., SENECA, A. & COX, C. J. 2014. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. *Genome Biol Evol,* 6**,** 897-911.

CLARKE, J. T., WARNOCK, R. C. & DONOGHUE, P. C. 2011. Establishing a time-scale for plant evolution. *New Phytol,* 192**,** 266-301.

CLOUGH, S. J. & BENT, A. F. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant Journal,* 16**,** 735-743.

COLOVOS, C. & YEATES, T. O. 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci,* 2**,** 1511-9.

CONAWAY, R. C. & CONAWAY, J. W. 1993. General initiation factors for RNA polymerase II. *Annu Rev Biochem,* 62**,** 161-90.

CONE, K. C., COCCIOLONE, S. M., BURR, F. A. & BURR, B. 1993. Maize Anthocyanin Regulatory Gene Pl Is a Duplicate of C1 That Functions in the Plant. *Plant Cell,* 5**,** 1795-1805.

COVE, D. 2005. The moss Physcomitrella patens. *Annual Review of Genetics,* 39**,** 339-358.

COVE, D., BEZANILLA, M., HARRIES, P. & QUATRANO, R. 2006. Mosses as model systems for the study of metabolism and development. *Annual Review of Plant Biology,* 57**,** 497-520.

COVE, D. J., KNIGHT, C. D. & LAMPARTER, T. 1997. Mosses as model systems. *Trends in Plant Science,* 2**,** 99-105.

COX, C. J., LI, B., FOSTER, P. G., EMBLEY, T. M. & CIVAN, P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Syst Biol,* 63**,** 272-9.

CRAMER, P., BUSHNELL, D. A. & KORNBERG, R. D. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom ngstrom resolution. *Science,* 292**,** 1863-1876.

CRONQUIST, A. 1988. *The evolution and classification of flowering plants,* Bronx, NY, New York Botanical Garden.

CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res,* 14**,** 1188-90.

CUI, L., WALL, P. K., LEEBENS-MACK, J. H., LINDSAY, B. G., SOLTIS, D. E., DOYLE, J. J., SOLTIS, P. S., CARLSON, J. E., ARUMUGANATHAN, K., BARAKAT, A., ALBERT, V. A., MA, H. & DEPAMPHILIS, C. W. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res,* 16**,** 738-49.

CUI, X., WANG, Q. D., YIN, W. Z., XU, H. Y., WILSON, Z. A., WEI, C. C., PAN, S. Y. & ZHANG, D. B. 2012. PMRD: a curated database for genes and mutants involved in plant male reproduction. *Bmc Plant Biology,* 12.

CZARAN, T. L. & HOEKSTRA, R. F. 2004. Evolution of sexual asymmetry. *Bmc Evolutionary Biology,* 4.

DASZKOWSKA-GOLEC, A., WOJNAR, W., ROSIKIEWICZ, M., SZAREJKO, I., MALUSZYNSKI, M., SZWEYKOWSKA-KULINSKA, Z. & JARMOLOWSKI, A. 2013. Arabidopsis suppressor

mutant of abh1 shows a new face of the already known players: ABH1 (CBP80) and ABI4-in response to ABA and abiotic stresses during seed germination. *Plant Mol Biol,* 81**,** 189-209.

DAVIDSON, R. M., HANSEY, C. N., GOWDA, M., CHILDS, K. L., LIN, H., VAILLANCOURT, B., SEKHON, R. S., DE LEON, N., KAEPPLER, S. M., JIANG, N. & BUELL, C. R. 2011. Utility of RNA Sequencing for Analysis of Maize Reproductive Transcriptomes. *The Plant Genome Journal,* 4**,** 191.

DAVIS, C. C., XI, Z. & MATHEWS, S. 2014. Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biol,* 12**,** 11.

DAVIS, E., CAIMENT, F., TORDOIR, X., CAVAILLE, J., FERGUSON-SMITH, A., COCKETT, N., GEORGES, M. & CHARLIER, C. 2005. RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Curr Biol,* 15**,** 743-9.

DAXINGER, L., HUNTER, B., SHEIKH, M., JAUVION, V., GASCIOLLI, V., VAUCHERET, H., MATZKE, M. & FURNER, I. 2008. Unexpected silencing effects from T-DNA tags in Arabidopsis. *Trends Plant Sci,* 13**,** 4-6.

DE BODT, S., MAERE, S. & VAN DE PEER, Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol,* 20**,** 591-7.

DELANO, W. L. 2004. Use of PYMOL as a communications tool for molecular science. *Abstracts of Papers of the American Chemical Society,* 228**,** U313-U314.

DENEKAMP, M. 2003. Integration of Wounding and Osmotic Stress Signals Determines the Expression of the AtMYB102 Transcription Factor Gene. *Plant Physiology,* 132**,** 1415-1423.

DODDRELL, R. D., DUN, X. P., MOATE, R. M., JESSEN, K. R., MIRSKY, R. & PARKINSON, D. B. 2012. Regulation of Schwann cell differentiation and proliferation by the Pax-3 transcription factor. *Glia,* 60**,** 1269-78.

DRUMMOND, A. J., HO, S. Y. W., PHILLIPS, M. J. & RAMBAUT, A. 2006. Relaxed phylogenetics and dating with confidence. *Plos Biology,* 4**,** 699-710.

DRYSDALE, C. M., DUENAS, E., JACKSON, B. M., REUSSER, U., BRAUS, G. H. & HINNEBUSCH, A. G. 1995. The Transcriptional Activator Gcn4 Contains Multiple Activation Domains That Are Critically Dependent on Hydrophobic Amino-Acids. *Molecular and Cellular Biology,* 15**,** 1220-1233.

DUBOS, C., STRACKE, R., GROTEWOLD, E., WEISSHAAR, B., MARTIN, C. & LEPINIEC, L. 2010. MYB transcription factors in Arabidopsis. *Trends Plant Sci,* 15**,** 573-81.

DURBARRY, A., VIZIR, I. & TWELL, D. 2005. Male germ line development in Arabidopsis. duo pollen mutants reveal gametophytic regulators of generative cell cycle progression. *Plant Physiol,* 137**,** 297-307.

DYER, B. W., FERRER, F. A., KLINEDINST, D. K. & RODRIGUEZ, R. 2000. A noncommercial dual luciferase enzyme assay system for reporter gene analysis. *Anal Biochem,* 282**,** 158-61.

EDGER, P. P. & PIRES, J. C. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res,* 17**,** 699-717.

EDWARDS, S. V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution,* 63**,** 1-19.

ELFVING, N., DAVOINE, C., BENLLOCH, R., BLOMBERG, J., BRANNSTROM, K., MULLER, D., NILSSON, A., ULFSTEDT, M., RONNE, H., WINGSLE, G., NILSSON, O. & BJORKLUND, S. 2011. The Arabidopsis thaliana Med25 mediator subunit integrates environmental cues to control plant development. *Proc Natl Acad Sci U S A,* 108**,** 8245-50.

ELLEGREN, H. & PARSCH, J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet,* 8**,** 689-98.

EMERY, J. F., FLOYD, S. K., ALVAREZ, J., ESHED, Y., HAWKER, N. P., IZHAKI, A., BAUM, S. F. & BOWMAN, J. L. 2003. Radial Patterning of Arabidopsis Shoots by Class III HD-ZIP and KANADI Genes. *Current Biology,* 13**,** 1768-1774.

ENRIGHT, A. J., JOHN, B., GAUL, U., TUSCHL, T., SANDER, C. & MARKS, D. S. 2003. MicroRNA targets in Drosophila. *Genome Biol,* 5**,** R1.

ERKINA, T. Y. & ERKINE, A. M. 2016. Nucleosome distortion as a possible mechanism of transcription activation domain function. *Epigenetics Chromatin,* 9**,** 40.

EYRE-WALKER, A. & KEIGHTLEY, P. D. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution,* 26**,** 2097-2108.

FASOLI, M., DAL SANTO, S., ZENONI, S., TORNIELLI, G. B., FARINA, L., ZAMBONI, A., PORCEDDU, A., VENTURINI, L., BICEGO, M., MURINO, V., FERRARINI, A., DELLEDONNE, M. & PEZZOTTI, M. 2012. The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell,* 24**,** 3489-505.

FAWCETT, J. A., MAERE, S. & VAN DE PEER, Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A,* 106**,** 5737-42.

FELDMAN, M. & KISLEV, M. E. 2007. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. *Israel Journal of Plant Sciences,* 55**,** 207-221.

FELLER, A., MACHEMER, K., BRAUN, E. L. & GROTEWOLD, E. 2011. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *Plant J,* 66**,** 94-116.

FERGUSON, A. C., PEARCE, S., BAND, L. R., YANG, C. Y., FERJENTSIKOVA, I., KING, J., YUAN, Z., ZHANG, D. B. & WILSON, Z. A. 2017. Biphasic regulation of the transcription factor ABORTED MICROSPORES (AMS) is essential for tapetum and pollen development in Arabidopsis. *New Phytologist,* 213**,** 778-790.

FERRIS, P. J., PAVLOVIC, C., FABRY, S. & GOODENOUGH, U. W. 1997. Rapid evolution of sex-related genes in Chlamydomonas. *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 8634-8639.

FERRIS, P. J., WOESSNER, J. P. & GOODENOUGH, U. W. 1996. A sex recognition glycoprotein is encoded by the plus mating-type gene fus1 of Chlamydomonas reinhardtii. *Molecular Biology of the Cell,* 7**,** 1235-1248.

FINET, C., TIMME, RUTH E., DELWICHE, CHARLES F. & MARLÉTAZ, F. 2012. Multigene Phylogeny of the Green Lineage Reveals the Origin and Diversification of Land Plants. *Current Biology,* 22**,** 1456-1457.

FOSTER, C. S. P., SAUQUET, H., VAN DER MERWE, M., MCPHERSON, H., ROSSETTO, M. & HO, S. Y. W. 2017. Evaluating the Impact of Genomic Data and Priors on Bayesian Estimates of the Angiosperm Evolutionary Timescale. *Syst Biol,* 66**,** 338-351.

FRANCO-ZORRILLA, J. M., LOPEZ-VIDRIERO, I., CARRASCO, J. L., GODOY, M., VERA, P. & SOLANO, R. 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci U S A,* 111**,** 2367-72.

GALWAY, M. E. & HARDHAM, A. R. 1991. Immunofluorescent Localization of Microtubules Throughout the Cell-Cycle in the Green-Alga Mougeotia (Zygnemataceae). *American Journal of Botany,* 78**,** 451-461.

GARCIA, D. 2008. A miRacle in plant development: role of microRNAs in cell differentiation and patterning. *Semin Cell Dev Biol,* 19**,** 586-95.

GAUT, B. S. & DOEBLEY, J. F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci U S A,* 94**,** 6809-14.

GEHRING, W. J. 1992. The homeobox in perspective. *Trends Biochem Sci,* 17**,** 277-80.

GHODKE, H., WANG, H., HSIEH, C. L., WOLDEMESKEL, S., WATKINS, S. C., RAPIC-OTRIN, V. & VAN HOUTEN, B. 2014. Single-molecule analysis reveals human UV-damaged DNA-binding protein (UV-DDB) dimerizes on DNA via multiple kinetic intermediates. *Proc Natl Acad Sci U S A,* 111**,** E1862-71.

GILL, G. 2001. Regulation of the initiation of eukaryotic transcription. *Essays Biochem,* 37**,** 33-43.

GILL, G. & PTASHNE, M. 1987. Mutants of GAL4 protein altered in an activation function. *Cell,* 51**,** 121-6.

GOFFINET, B. & SHAW, A. J. 2009. *Bryophyte biology,* Cambridge, Cambridge University Press.

GOMEZ, J. F., TALLE, B. & WILSON, Z. A. 2015. Anther and pollen development: A conserved developmental pathway. *J Integr Plant Biol,* 57**,** 876-91.

GONTCHAROV, A. A., MARIN, B. A. & MELKONIAN, M. A. 2003. Molecular phylogeny of conjugating green algae (Zygnemophyceae, Streptophyta) inferred from SSU rDNA sequence comparisons. *J Mol Evol,* 56**,** 89-104.

GORELICK, R. & HENG, H. H. Q. 2011. Sex Reduces Genetic Variation: A Multidisciplinary Review. *Evolution,* 65**,** 1088-1098.

GRAHAM, L. E., COOK, M. E. & BUSSE, J. S. 2000. The origin of plants: body plan changes contributing to a major evolutionary radiation. *Proc Natl Acad Sci U S A,* 97**,** 4535-40.

GRANT-DOWNTON, R., KOURMPETLI, S., HAFIDH, S., KHATAB, H., LE TRIONNAIRE, G., DICKINSON, H. & TWELL, D. 2013. Artificial microRNAs reveal cell-specific differences in small RNA activity in pollen. *Curr Biol,* 23**,** R599-601.

GRAY, R. D., DRUMMOND, A. J. & GREENHILL, S. J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science,* 323**,** 479-83.

GREEN, R. E., KRAUSE, J., BRIGGS, A. W., MARICIC, T., STENZEL, U., KIRCHER, M., PATTERSON, N., LI, H., ZHAI, W., FRITZ, M. H., HANSEN, N. F., DURAND, E. Y., MALASPINAS, A. S., JENSEN, J. D., MARQUES-BONET, T., ALKAN, C., PRUFER, K., MEYER, M., BURBANO, H. A., GOOD, J. M., SCHULTZ, R., AXIMU-PETRI, A., BUTTHOF, A., HOBER, B., HOFFNER, B., SIEGEMUND, M., WEIHMANN, A., NUSBAUM, C., LANDER, E. S., RUSS, C., NOVOD, N., AFFOURTIT, J., EGHOLM, M., VERNA, C., RUDAN, P., BRAJKOVIC, D., KUCAN, Z., GUSIC, I., DORONICHEV, V. B., GOLOVANOVA, L. V., LALUEZA-FOX, C., DE LA RASILLA, M., FORTEA, J., ROSAS, A., SCHMITZ, R. W., JOHNSON, P. L., EICHLER, E. E., FALUSH, D., BIRNEY, E., MULLIKIN, J. C., SLATKIN, M., NIELSEN, R., KELSO, J., LACHMANN, M., REICH, D. & PAABO, S. 2010. A draft sequence of the Neandertal genome. *Science,* 328**,** 710-22.

GRENFELL, B. T., PYBUS, O. G., GOG, J. R., WOOD, J. L., DALY, J. M., MUMFORD, J. A. & HOLMES, E. C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science,* 303**,** 327-32.

GRIFFITHS-JONES, S., GROCOCK, R. J., VAN DONGEN, S., BATEMAN, A. & ENRIGHT, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research,* 34**,** D140-D144.

GRIFFITHS-JONES, S., SAINI, H. K., VAN DONGEN, S. & ENRIGHT, A. J. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res,* 36**,** D154-8.

GRONAU, I., HUBISZ, M. J., GULKO, B., DANKO, C. G. & SIEPEL, A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet,* 43**,** 1031-4.

GROTH-MALONEK, M., PRUCHNER, D., GREWE, F. & KNOOP, V. 2005. Ancestors of trans-splicing mitochondrial introns support serial sister group relationships of hornworts and mosses with vascular plants. *Mol Biol Evol,* 22**,** 117-25.

GU, J. N., ZHU, J., YU, Y., TENG, X. D., LOU, Y., XU, X. F., LIU, J. L. & YANG, Z. N. 2014. DYT1 directly regulates the expression of TDF1 for tapetum development and pollen wall formation in Arabidopsis. *Plant J,* 80**,** 1005-13.

GU, W. & ROEDER, R. G. 1997. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell,* 90**,** 595-606.

GUBLER, F., KALLA, R., ROBERTS, J. K. & JACOBSEN, J. V. 1995. Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter. *Plant Cell,* 7**,** 1879-91.

GUBLER, F., RAVENTOS, D., KEYS, M., WATTS, R., MUNDY, J. & JACOBSEN, J. V. 1999. Target genes and regulatory domains of the GAMYB transcriptional activator in cereal aleurone. *Plant J,* 17**,** 1-9.

HAFIDH, S., BREZNENOVA, K., RUZICKA, P., FECIKOVA, J., CAPKOVA, V. & HONYS, D. 2012. Comprehensive analysis of tobacco pollen transcriptome unveils common pathways in polar cell expansion and underlying heterochronic shift during spermatogenesis. *BMC Plant Biol,* 12**,** 24.

HAGA, N., KATO, K., MURASE, M., ARAKI, S., KUBO, M., DEMURA, T., SUZUKI, K., MULLER, I., VOSS, U., JURGENS, G. & ITO, M. 2007. R1R2R3-Myb proteins positively regulate cytokinesis through activation of KNOLLE transcription in Arabidopsis thaliana. *Development,* 134**,** 1101-10.

HAMILTON, W. D., AXELROD, R. & TANESE, R. 1990. Sexual Reproduction as an Adaptation to Resist Parasites (a Review). *Proceedings of the National Academy of Sciences of the United States of America,* 87**,** 3566-3573.

HANAHAN, D. 1983. Studies on Transformation of Escherichia-Coli with Plasmids. *Journal of Molecular Biology,* 166**,** 557-580.

HANANO, S. & GOTO, K. 2011. Arabidopsis TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression. *Plant Cell,* 23**,** 3172-84.

HANIN, M. & PASZKOWSKI, J. 2003. Plant genome modification by homologous recombination. *Current Opinion in Plant Biology,* 6**,** 157-162.

HE, L. & HANNON, G. J. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet,* 5**,** 522-31.

HEDGES, S. B. 2002. The origin and evolution of model organisms. *Nat Rev Genet,* 3**,** 838-49.

HEFFERNAN, R., DEHZANGI, A., LYONS, J., PALIWAL, K., SHARMA, A., WANG, J., SATTAR, A., ZHOU, Y. & YANG, Y. 2016. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics,* 32**,** 843-9.

HEIM, M. A., JAKOBY, M., WERBER, M., MARTIN, C., WEISSHAAR, B. & BAILEY, P. C. 2003. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol Biol Evol,* 20**,** 735-47.

HEMM, M. R., HERRMANN, K. M. & CHAPPLE, C. 2001. AtMYB4: a transcription factor general in the battle against UV. *Trends Plant Sci,* 6**,** 135-6.

HERRERA, F. J. & TRIEZENBERG, S. J. 2004. VP16-dependent association of chromatin-modifying coactivators and underrepresentation of histones at immediate-early gene promoters during herpes simplex virus infection. *Journal of Virology,* 78**,** 9689-9696.

HIGGINS, D. G. & SHARP, P. M. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene,* 73**,** 237-44.

HIGGINSON, T., LI, S. F. & PARISH, R. W. 2003. AtMYB103 regulates tapetum and trichome development in Arabidopsis thaliana. *Plant J,* 35**,** 177-92.

HIGO, A., NIWA, M., YAMATO, K. T., YAMADA, L., SAWADA, H., SAKAMOTO, T., KURATA, T., SHIRAKAWA, M., ENDO, M., SHIGENOBU, S., YAMAGUCHI, K., ISHIZAKI, K., NISHIHAMA, R., KOHCHI, T. & ARAKI, T. 2016. Transcriptional Framework of Male Gametogenesis in the Liverwort Marchantia polymorpha L. *Plant Cell Physiol,* 57**,** 325-38.

HOLLENDER, C. A., KANG, C., DARWISH, O., GERETZ, A., MATTHEWS, B. F., SLOVIN, J., ALKHAROUF, N. & LIU, Z. 2014. Floral Transcriptomes in Woodland Strawberry Uncover Developing Receptacle and Anther Gene Networks. *Plant Physiol,* 165**,** 1062-1075.

HOLTHAUSEN, J. T., WYMAN, C. & KANAAR, R. 2010. Regulation of DNA strand exchange in homologous recombination. *DNA Repair (Amst),* 9**,** 1264-72.

HONYS, D. & TWELL, D. 2003. Comparative analysis of the Arabidopsis pollen transcriptome. *Plant Physiol,* 132**,** 640-52.

HONYS, D. & TWELL, D. 2004. Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biology,* 5.

HOPE, I. A., MAHADEVAN, S. & STRUHL, K. 1988. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature,* 333**,** 635-40.

HOPE, I. A. & STRUHL, K. 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell,* 46**,** 885-94.

HOROWITZ, J. M. & ENGLAND, J. L. 2017. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc Natl Acad Sci U S A,* 114**,** 7565-7570.

HSU, P. D., LANDER, E. S. & ZHANG, F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell,* 157**,** 1262-78.

HU, S. S., DILCHER, D. L., JARZEN, D. M. & TAYLOR, D. W. 2008. Early steps of angiosperm-pollinator coevolution. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 240-245.

HUA, T., VEMURI, K., NIKAS, S. P., LAPRAIRIE, R. B., WU, Y., QU, L., PU, M., KORDE, A., JIANG, S., HO, J. H., HAN, G. W., DING, K., LI, X., LIU, H., HANSON, M. A., ZHAO, S., BOHN, L. M., MAKRIYANNIS, A., STEVENS, R. C. & LIU, Z. J. 2017. Crystal structures of agonist-bound human cannabinoid receptor CB1. *Nature*.

HUA, T., VEMURI, K., PU, M., QU, L., HAN, G. W., WU, Y., ZHAO, S., SHUI, W., LI, S., KORDE, A., LAPRAIRIE, R. B., STAHL, E. L., HO, J. H., ZVONOK, N., ZHOU, H., KUFAREVA, I., WU, B., ZHAO, Q., HANSON, M. A., BOHN, L. M., MAKRIYANNIS, A., STEVENS, R. C. & LIU, Z. J. 2016. Crystal Structure of the Human Cannabinoid Receptor CB1. *Cell,* 167**,** 750-762 e14.

HUFFMAN, J. L. & BRENNAN, R. G. 2002. Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Current Opinion in Structural Biology,* 12**,** 98-106.

HURST, L. D. 1996. Why are there only two sexes? *Proceedings of the Royal Society B-Biological Sciences,* 263**,** 415-422.

IKEDA, M., MITSUDA, N. & OHME-TAKAGI, M. 2009. Arabidopsis WUSCHEL is a bifunctional transcription factor that acts as a repressor in stem cell regulation and as an activator in floral patterning. *Plant Cell,* 21**,** 3493-505.

INTERNATIONAL WHEAT GENOME SEQUENCING, C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science,* 345**,** 1251788.

INZE, D. & DE VEYLDER, L. 2006. Cell cycle regulation in plant development. *Annu Rev Genet,* 40**,** 77-105.

ISHIZAKI, K., CHIYODA, S., YAMATO, K. T. & KOHCHI, T. 2008. Agrobacterium-Mediated Transformation of the Haploid Liverwort Marchantia polymorpha L., an Emerging Model for Plant Biology. *Plant and Cell Physiology,* 49**,** 1084-1091.

ISHIZAKI, K., JOHZUKA-HISATOMI, Y., ISHIDA, S., IIDA, S. & KOHCHI, T. 2013. Homologous recombination-mediated gene targeting in the liverwort Marchantia polymorpha L. *Sci Rep,* 3**,** 1532.

ITO, M. 2005. Conservation and diversification of three-repeat Myb transcription factors in plants. *J Plant Res,* 118**,** 61-9.

ITO, M., ARAKI, S., MATSUNAGA, S., ITOH, T., NISHIHAMA, R., MACHIDA, Y., DOONAN, J. H. & WATANABE, A. 2001. G2/M-phase-specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors. *Plant Cell,* 13**,** 1891-905.

ITO, M., IWASE, M., KODAMA, H., LAVISSE, P., KOMAMINE, A., NISHIHAMA, R., MACHIDA, Y. & WATANABE, A. 1998. A novel cis-acting element in promoters of plant B-type cyclin genes activates M phase-specific transcription. *Plant Cell,* 10**,** 331-41.

ITO, T., NAGATA, N., YOSHIBA, Y., OHME-TAKAGI, M., MA, H. & SHINOZAKI, K. 2007. Arabidopsis MALE STERILITY1 encodes a PHD-type transcription factor and regulates pollen and tapetum development. *Plant Cell,* 19**,** 3549-62.

ITOH, T., FAIRALL, L., MUSKETT, F. W., MILANO, C. P., WATSON, P. J., ARNAUDO, N., SALEH, A., MILLARD, C. J., EL-MEZGUELDI, M., MARTINO, F. & SCHWABE, J. W. 2015. Structural and functional characterization of a cell cycle associated HDAC1/2 complex reveals the structural basis for complex assembly and nucleosome targeting. *Nucleic Acids Res,* 43**,** 2033-44.

IWAKAWA, H., SHINMYO, A. & SEKINE, M. 2006. Arabidopsis CDKA;1, a cdc2 homologue, controls proliferation of generative cells in male gametogenesis. *Plant J,* 45**,** 819-31.

JAILLON, O., AURY, J. M., NOEL, B., POLICRITI, A., CLEPET, C., CASAGRANDE, A., CHOISNE, N., AUBOURG, S., VITULO, N., JUBIN, C., VEZZI, A., LEGEAI, F., HUGUENEY, P., DASILVA, C., HORNER, D., MICA, E., JUBLOT, D., POULAIN, J., BRUYERE, C., BILLAULT, A., SEGURENS, B., GOUYVENOUX, M., UGARTE, E., CATTONARO, F., ANTHOUARD, V., VICO, V., DEL FABBRO, C., ALAUX, M., DI GASPERO, G., DUMAS, V., FELICE, N., PAILLARD, S., JUMAN, I., MOROLDO, M., SCALABRIN, S., CANAGUIER, A., LE CLAINCHE, I., MALACRIDA, G., DURAND, E., PESOLE, G., LAUCOU, V., CHATELET, P., MERDINOGLU, D., DELLEDONNE, M., PEZZOTTI, M., LECHARNY, A., SCARPELLI, C., ARTIGUENAVE, F., PE, M. E., VALLE, G., MORGANTE, M., CABOCHE, M., ADAM-BLONDON, A. F., WEISSENBACH, J., QUETIER, F., WINCKER, P. & FRENCH-ITALIAN PUBLIC CONSORTIUM FOR GRAPEVINE GENOME, C. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature,* 449**,** 463-7.

JANG, G., YI, K., PIRES, N. D., MENAND, B. & DOLAN, L. 2011. RSL genes are sufficient for rhizoid system development in early diverging land plants. *Development,* 138**,** 2273-81.

JANSEN, R. K., CAI, Z., RAUBESON, L. A., DANIELL, H., DEPAMPHILIS, C. W., LEEBENS-MACK, J., MULLER, K. F., GUISINGER-BELLIAN, M., HABERLE, R. C., HANSEN, A. K., CHUMLEY, T. W., LEE, S. B., PEERY, R., MCNEAL, J. R., KUEHL, J. V. & BOORE, J. L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A,* 104**,** 19369-74.

JENKINS, G. I. & COVE, D. J. 1983. Light Requirements for Regeneration of Protoplasts of the Moss Physcomitrella-Patens. *Planta,* 157**,** 39-45.

JIANG, C., GU, X. & PETERSON, T. 2004. Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of Arabidopsis and Oryza sativa L. ssp. indica. *Genome Biol,* 5**,** R46.

JIAO, Y., WICKETT, N. J., AYYAMPALAYAM, S., CHANDERBALI, A. S., LANDHERR, L., RALPH, P. E., TOMSHO, L. P., HU, Y., LIANG, H., SOLTIS, P. S., SOLTIS, D. E., CLIFTON, S. W., SCHLARBAUM, S. E., SCHUSTER, S. C., MA, H., LEEBENS-MACK, J. & DEPAMPHILIS, C. W. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature,* 473**,** 97-100.

JIN, H. L. & MARTIN, C. 1999. Multifunctionality and diversity within the plant MYB-gene family. *Plant Molecular Biology,* 41**,** 577-585.

JOAQUIN, M. & WATSON, R. J. 2003. Cell cycle regulation by the B-Myb transcription factor. *Cell Mol Life Sci,* 60**,** 2389-401.

JOHN, B., ENRIGHT, A. J., ARAVIN, A., TUSCHL, T., SANDER, C. & MARKS, D. S. 2004. Human MicroRNA targets. *PLoS Biol,* 2**,** e363.

JOHNSTON, R. J. & HOBERT, O. 2003. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature,* 426**,** 845-9.

KACHMAN, T., OWEN, J. A. & ENGLAND, J. L. 2017. Self-Organized Resonance during Search of a Diverse Chemical Space. *Phys Rev Lett,* 119**,** 038001.

KAKIDANI, H. & PTASHNE, M. 1988. GAL4 activates gene expression in mammalian cells. *Cell,* 52**,** 161-7.

KAMISUGI, Y., SCHAEFER, D. G., KOZAK, J., CHARLOT, F., VRIELYNCK, N., HOLA, M., ANGELIS, K. J., CUMING, A. C. & NOGUE, F. 2012. MRE11 and RAD50, but not NBS1, are essential for gene targeting in the moss Physcomitrella patens. *Nucleic Acids Res,* 40**,** 3496-510.

KAMMERER, W. & COVE, D. J. 1996. Genetic analysis of the effects of re-transformation of transgenic lines of the moss Physcomitrella patens. *Molecular & General Genetics,* 250**,** 380-382.

KAPRAUN, D. F. 2007. Nuclear DNA content estimates in green algal lineages: chlorophyta and streptophyta. *Ann Bot,* 99**,** 677-701.

KARIMI, M., INZE, D. & DEPICKER, A. 2002. GATEWAY vectors for Agrobacterium-mediated plant transformation. *Trends Plant Sci,* 7**,** 193-5.

KARIN, M. 1990. Too many transcription factors: positive and negative interactions. *New Biol,* 2**,** 126-31.

KAROL, K. G., ARUMUGANATHAN, K., BOORE, J. L., DUFFY, A. M., EVERETT, K. D., HALL, J. D., HANSEN, S. K., KUEHL, J. V., MANDOLI, D. F., MISHLER, B. D., OLMSTEAD, R. G., RENZAGLIA, K. S. & WOLF, P. G. 2010. Complete plastome sequences of Equisetum arvense and Isoetes flaccida: implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evol Biol,* 10**,** 321.

KAROL, K. G., MCCOURT, R. M., CIMINO, M. T. & DELWICHE, C. F. 2001. The closest living relatives of land plants. *Science,* 294**,** 2351-3.

KASAHARA, R. D., PORTEREIKO, M. F., SANDAKLIE-NIKOLOVA, L., RABIGER, D. S. & DREWS, G. N. 2005. MYB98 is required for pollen tube guidance and synergid cell differentiation in Arabidopsis. *Plant Cell,* 17**,** 2981-92.

KASSAHN, K. S., DANG, V. T., WILKINS, S. J., PERKINS, A. C. & RAGAN, M. A. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res,* 19**,** 1404-18.

KAUL, S., KOO, H. L., JENKINS, J., RIZZO, M., ROONEY, T., TALLON, L. J., FELDBLYUM, T., NIERMAN, W., BENITO, M. I., LIN, X. Y., TOWN, C. D., VENTER, J. C., FRASER, C. M., TABATA, S., NAKAMURA, Y., KANEKO, T., SATO, S., ASAMIZU, E., KATO, T., KOTANI, H., SASAMOTO, S., ECKER, J. R., THEOLOGIS, A., FEDERSPIEL, N. A., PALM, C. J., OSBORNE, B. I., SHINN, P., CONWAY, A. B., VYSOTSKAIA, V. S., DEWAR, K., CONN, L., LENZ, C. A., KIM, C. J., HANSEN, N. F., LIU, S. X., BUEHLER, E., ALTAFI, H., SAKANO, H., DUNN, P., LAM, B., PHAM, P. K., CHAO, Q., NGUYEN, M., YU, G. X., CHEN, H. M., SOUTHWICK, A., LEE, J. M., MIRANDA, M., TORIUMI, M. J., DAVIS, R. W., WAMBUTT, R., MURPHY, G., DUSTERHOFT, A., STIEKEMA, W., POHL, T., ENTIAN, K. D., TERRYN, N., VOLCKAERT, G., SALANOUBAT, M., CHOISNE, N., RIEGER, M., ANSORGE, W., UNSELD, M., FARTMANN, B., VALLE, G., ARTIGUENAVE, F., WEISSENBACH, J., QUETIER, F., WILSON, R. K., DE LA BASTIDE, M., SEKHON, M., HUANG, E., SPIEGEL, L., GNOJ, L., PEPIN, K., MURRAY, J., JOHNSON, D., HABERMANN, K., DEDHIA, N., PARNELL, L., PRESTON, R., HILLIER, L., CHEN, E., MARRA, M., MARTIENSSEN, R., MCCOMBIE, W. R., MAYER, K., WHITE, O., BEVAN, M., LEMCKE, K., CREASY, T. H., BIELKE, C., HAAS, B., HAASE, D., MAITI, R., RUDD, S., PETERSON, J., SCHOOF, H., FRISHMAN, D., MORGENSTERN, B., et al. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* 408**,** 796-815.

KEIGHTLEY, P. D. & EYRE-WALKER, A. 2012. Estimating the Rate of Adaptive Molecular Evolution When the Evolutionary Divergence Between Species is Small. *Journal of Molecular Evolution,* 74**,** 61-68.

KELLIS, M., BIRREN, B. W. & LANDER, E. S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature,* 428**,** 617-24.

KELLIS, M., PATTERSON, N., ENDRIZZI, M., BIRREN, B. & LANDER, E. S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* 423**,** 241-54.

KENRICK, P. & CRANE, P. R. 1997. The origin and early evolution of plants on land. *Nature,* 389**,** 33-39.

KENRICK, P., WELLMAN, C. H., SCHNEIDER, H. & EDGECOMBE, G. D. 2012. A timeline for terrestrialization: consequences for the carbon cycle in the Palaeozoic. *Philos Trans R Soc Lond B Biol Sci,* 367**,** 519-36.

KIEFER, F., ARNOLD, K., KUNZLI, M., BORDOLI, L. & SCHWEDE, T. 2009. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res,* 37**,** D387-92.

KIM, H. J., OH, S. A., BROWNFIELD, L., HONG, S. H., RYU, H., HWANG, I., TWELL, D. & NAM, H. G. 2008a. Control of plant germline proliferation by SCF(FBL17) degradation of cell cycle inhibitors. *Nature,* 455**,** 1134-7.

KIM, S., YANG, J. Y., XU, J., JANG, I. C., PRIGGE, M. J. & CHUA, N. H. 2008b. Two cap-binding proteins CBP20 and CBP80 are involved in processing primary MicroRNAs. *Plant Cell Physiol,* 49**,** 1634-44.

KIM, V. N. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol,* 6**,** 376-85.

KIM, Y. J., KIM, J. H., LEE, C. E., MOK, Y. G., CHOI, J. S., SHIN, H. S. & HWANG, S. 2006. Expression of yeast transcriptional activator MSN1 promotes accumulation of chromium and sulfur by enhancing sulfate transporter level in plants. *FEBS Lett,* 580**,** 206-10.

KIRIAKIDOU, M., NELSON, P. T., KOURANOV, A., FITZIEV, P., BOUYIOUKOS, C., MOURELATOS, Z. & HATZIGEORGIOU, A. 2004. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev,* 18**,** 1165-78.

KIRIK, V., LEE, M. M., WESTER, K., HERRMANN, U., ZHENG, Z., OPPENHEIMER, D., SCHIEFELBEIN, J. & HULSKAMP, M. 2005. Functional diversification of MYB23 and GL1 genes in trichome morphogenesis and initiation. *Development,* 132**,** 1477-85.

KLEMPNAUER, K. H., GONDA, T. J. & BISHOP, J. M. 1982. Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. *Cell,* 31**,** 453-63.

KLUG, A. 2010. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem,* 79**,** 213-31.

KLUG, A. & RHODES, D. 1987. Zinc fingers: a novel protein fold for nucleic acid recognition. *Cold Spring Harb Symp Quant Biol,* 52**,** 473-82.

KORNBERG, R. D. 2007. The molecular basis of eukaryotic transcription. *Proc Natl Acad Sci U S A,* 104**,** 12955-61.

KOZOMARA, A. & GRIFFITHS-JONES, S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research,* 42**,** D68-D73.

KROGH, B. O. & SYMINGTON, L. S. 2004. Recombination proteins in yeast. *Annu Rev Genet,* 38**,** 233-71.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

KYO, M., HATTORI, S., YAMAJI, N., PECHAN, P. & FUKUI, H. 2003. Cloning and characterization of cDNAs associated with the embryogenic dedifferentiation of tobacco immature pollen grains. *Plant Science,* 164**,** 1057-1066.

KYTE, J. & DOOLITTLE, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol,* 157**,** 105-32.

LABBY, K. J. 2013. Protein visualization workshop: Interactive biochemistry using PyMOL and the PDB. *Abstracts of Papers of the American Chemical Society,* 246.

LAGOS-QUINTANA, M., RAUHUT, R., LENDECKEL, W. & TUSCHL, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science,* 294**,** 853-8.

LAITY, J. H., LEE, B. M. & WRIGHT, P. E. 2001. Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology,* 11**,** 39-46.

LAM, E. W. & WATSON, R. J. 1993. An E2F-binding site mediates cell-cycle regulated repression of mouse B-myb transcription. *EMBO J,* 12**,** 2705-13.

LANDGRAF, P., RUSU, M., SHERIDAN, R., SEWER, A., IOVINO, N., ARAVIN, A., PFEFFER, S., RICE, A., KAMPHORST, A. O., LANDTHALER, M., LIN, C., SOCCI, N. D., HERMIDA, L., FULCI, V., CHIARETTI, S., FOÀ, R., SCHLIWKA, J., FUCHS, U., NOVOSEL, A., MÜLLER, R.-U., SCHERMER, B., BISSELS, U., INMAN, J., PHAN, Q., CHIEN, M., WEIR, D. B., CHOKSI, R., DE VITA, G., FREZZETTI, D., TROMPETER, H.-I., HORNUNG, V., TENG, G., HARTMANN, G., PALKOVITS, M., DI LAURO, R., WERNET, P., MACINO, G., ROGLER, C. E., NAGLE, J. W., JU, J., PAPAVASILIOU, F. N., BENZING, T., LICHTER, P., TAM, W., BROWNSTEIN, M. J., BOSIO, A., BORKHARDT, A., RUSSO, J. J., SANDER, C., ZAVOLAN, M. & TUSCHL, T. 2007. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell,* 129**,** 1401-1414.

LANG, D., LU, M. M., HUANG, L., ENGLEKA, K. A., ZHANG, M. Z., CHU, E. Y., LIPNER, S., SKOULTCHI, A., MILLAR, S. E. & EPSTEIN, J. A. 2005. Pax3 functions at a nodal point in melanocyte stem cell differentiation. *Nature,* 433**,** 884-887.

LANG, V., USADEL, B. & OBERMEYER, G. 2015. De novo sequencing and analysis of the lily pollen transcriptome: an open access data source for an orphan plant species. *Plant Mol Biol,* 87**,** 69-80.

LATCHMAN, D. S. 1997. Transcription factors: an overview. *Int J Biochem Cell Biol,* 29**,** 1305-12.

LAU, N. C., LIM, L. P., WEINSTEIN, E. G. & BARTEL, D. P. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science,* 294**,** 858-62.

LAURIN-LEMAY, S., BRINKMANN, H. & PHILIPPE, H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol,* 22**,** R593-4.

LEE, M. M. & SCHIEFELBEIN, J. 1999. WEREWOLF, a MYB-related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning. *Cell,* 99**,** 473-83.

LEE, R. C. & AMBROS, V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science,* 294**,** 862-4.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75**,** 843-54.

LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., HANNETT, N. M., HARBISON, C. T., THOMPSON, C. M., SIMON, I., ZEITLINGER, J., JENNINGS, E. G., MURRAY, H. L., GORDON, D. B., REN, B., WYRICK, J. J., TAGNE, J. B., VOLKERT, T. L., FRAENKEL, E., GIFFORD, D. K. & YOUNG, R. A. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science,* 298**,** 799-804.

LEE, T. I. & YOUNG, R. A. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet,* 34**,** 77-137.

LELIAERT, F., SMITH, D. R., MOREAU, H., HERRON, M. D., VERBRUGGEN, H., DELWICHE, C. F. & DE CLERCK, O. 2012. Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences,* 31**,** 1-46.

LEMIEUX, C., OTIS, C. & TURMEL, M. 2007. A clade uniting the green algae Mesostigma viride and Chlorokybus atmophyticus represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol,* 5**,** 2.

LEMON, B. & TJIAN, R. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev,* 14**,** 2551-69.

LEWIS, B. P., SHIH, I. H., JONES-RHOADES, M. W., BARTEL, D. P. & BURGE, C. B. 2003. Prediction of mammalian microRNA targets. *Cell,* 115**,** 787-98.

LEWIS, L. A. & MCCOURT, R. M. 2004. Green algae and the origin of land plants. *Am J Bot,* 91**,** 1535-56.

LEYDON, A. R., BEALE, K. M., WORONIECKA, K., CASTNER, E., CHEN, J., HORGAN, C., PALANIVELU, R. & JOHNSON, M. A. 2013. Three MYB transcription factors control pollen tube differentiation required for sperm release. *Curr Biol,* 23**,** 1209-14.

LI, H. & DURBIN, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature,* 475**,** 493-6.

LI, H., MENG, F., GUO, C., WANG, Y., XIE, X., ZHU, T., ZHOU, S., MA, H., SHAN, H. & KONG, H. 2014a. MeioBase: a comprehensive database for meiosis. *Front Plant Sci,* 5**,** 728.

LI, H., YUAN, Z., VIZCAY-BARRENA, G., YANG, C. Y., LIANG, W. Q., ZONG, J., WILSON, Z. A. & ZHANG, D. B. 2011a. PERSISTENT TAPETAL CELL1 Encodes a PHD-Finger Protein That Is Required for Tapetal Cell Death and Pollen Development in Rice. *Plant Physiology,* 156**,** 615-630.

LI, J., BLUE, R., ZEITLER, B., STRANGE, T. L., PEARL, J. R., HUIZINGA, D. H., EVANS, S., GREGORY, P. D., URNOV, F. D. & PETOLINO, J. F. 2013a. Activation domains for controlling plant gene expression using designed transcription factors. *Plant Biotechnol J,* 11**,** 671-80.

LI, J. & MILLAR, A. A. 2013. Expression of a microRNA-resistant target transgene misrepresents the functional significance of the endogenous microRNA: target gene relationship. *Mol Plant,* 6**,** 577-80.

LI, J., REICHEL, M. & MILLAR, A. A. 2014b. Determinants beyond both complementarity and cleavage govern microR159 efficacy in Arabidopsis. *PLoS Genet,* 10**,** e1004232.

LI, J., WU, Y. & QI, Y. 2014c. MicroRNAs in a multicellular green alga Volvox carteri. *Sci China Life Sci,* 57**,** 36-45.

LI, S. F., HIGGINSON, T. & PARISH, R. W. 1999. A novel MYB-related gene from Arabidopsis thaliana expressed in developing anthers. *Plant Cell Physiol,* 40**,** 343-7.

LI, X., BIAN, H., SONG, D., MA, S., HAN, N., WANG, J. & ZHU, M. 2013b. Flowering time control in ornamental gloxinia (Sinningia speciosa) by manipulation of miR159 expression. *Ann Bot,* 111**,** 791-9.

LI, Y., LI, C., DING, G. & JIN, Y. 2011b. Evolution of MIR159/319 microRNA genes and their post-transcriptional regulatory link to siRNA pathways. *BMC Evol Biol,* 11**,** 122.

LIANG, Y., TAN, Z. M., ZHU, L., NIU, Q. K., ZHOU, J. J., LI, M., CHEN, L. Q., ZHANG, X. Q. & YE, D. 2013. MYB97, MYB101 and MYB120 function as male factors that control pollen tube-synergid interaction in Arabidopsis thaliana fertilization. *PLoS Genet,* 9**,** e1003933.

LIN, P. C., LU, C. W., SHEN, B. N., LEE, G. Z., BOWMAN, J. L., ARTEAGA-VAZQUEZ, M. A., LIU, L. Y. D., HONG, S. F., LO, C. F., SU, G. M., KOHCHI, T., ISHIZAKI, K., ZACHGO, S., ALTHOFF, F., TAKENAKA, M., YAMATO, K. T. & LIN, S. S. 2016. Identification of miRNAs and Their Targets in the Liverwort Marchantia polymorpha by Integrating RNA-Seq and Degradome Analyses. *Plant and Cell Physiology,* 57**,** 339-358.

LIN, Z., YIN, K., ZHU, D., CHEN, Z., GU, H. & QU, L. J. 2007. AtCDC5 regulates the G2 to M transition of the cell cycle and is critical for the function of Arabidopsis shoot apical meristem. *Cell Res,* 17**,** 815-28.

LINDBLAD-TOH, K., GARBER, M., ZUK, O., LIN, M. F., PARKER, B. J., WASHIETL, S., KHERADPOUR, P., ERNST, J., JORDAN, G., MAUCELI, E., WARD, L. D., LOWE, C. B., HOLLOWAY, A. K., CLAMP, M., GNERRE, S., ALFOLDI, J., BEAL, K., CHANG, J., CLAWSON, H., CUFF, J., DI PALMA, F., FITZGERALD, S., FLICEK, P., GUTTMAN, M., HUBISZ, M. J., JAFFE, D. B., JUNGREIS, I., KENT, W. J., KOSTKA, D., LARA, M., MARTINS, A. L., MASSINGHAM, T., MOLTKE, I., RANEY, B. J., RASMUSSEN, M. D., ROBINSON, J., STARK, A., VILELLA, A. J., WEN, J., XIE, X., ZODY, M. C., BROAD INSTITUTE SEQUENCING, P., WHOLE GENOME ASSEMBLY, T., BALDWIN, J., BLOOM, T., CHIN, C. W., HEIMAN, D., NICOL, R., NUSBAUM, C., YOUNG, S., WILKINSON, J., WORLEY, K. C., KOVAR, C. L., MUZNY, D. M., GIBBS, R. A., BAYLOR COLLEGE OF MEDICINE HUMAN GENOME SEQUENCING CENTER SEQUENCING, T., CREE, A., DIHN, H. H., FOWLER, G., JHANGIANI, S., JOSHI, V., LEE, S., LEWIS, L. R., NAZARETH, L. V., OKWUONU, G., SANTIBANEZ, J., WARREN, W. C., MARDIS, E. R., WEINSTOCK, G. M., WILSON, R. K., GENOME INSTITUTE AT WASHINGTON, U., DELEHAUNTY, K., DOOLING, D., FRONIK, C., FULTON, L., FULTON, B., GRAVES, T., MINX, P., SODERGREN, E., BIRNEY, E., MARGULIES, E. H., HERRERO, J., GREEN, E. D., HAUSSLER, D., SIEPEL, A., GOLDMAN, N., POLLARD, K. S., PEDERSEN, J. S., LANDER, E. S. & KELLIS,

M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature,* 478**,** 476-82.

LINDNER, A. C., LANG, D., SEIFERT, M., PODLESAKOVA, K., NOVAK, O., STRNAD, M., RESKI, R. & VON SCHWARTZENBERG, K. 2014. Isopentenyltransferase-1 (IPT1) knockout in Physcomitrella together with phylogenetic analyses of IPTs provide insights into evolution of plant cytokinin biosynthesis. *J Exp Bot,* 65**,** 2533-43.

LIPINSKA, A., CORMIER, A., LUTHRINGER, R., PETERS, A. F., CORRE, E., GACHON, C. M., COCK, J. M. & COELHO, S. M. 2015. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga ectocarpus. *Mol Biol Evol,* 32**,** 1581-97.

LIPINSKA, A. P., D'HONDT, S., VAN DAMME, E. J. & DE CLERCK, O. 2013. Uncovering the genetic basis for early isogamete differentiation: a case study of Ectocarpus siliculosus. *BMC Genomics,* 14**,** 909.

LIPINSKA, A. P., VAN DAMME, E. J. & DE CLERCK, O. 2016. Molecular evolution of candidate male reproductive genes in the brown algal model Ectocarpus. *BMC Evol Biol,* 16**,** 5.

LIPSICK, J. S. 1996. One billion years of Myb. *Oncogene,* 13**,** 223-35.

LIU, Y., GONG, W., HUANG, C. C., HERR, W. & CHENG, X. 1999. Crystal structure of the conserved core of the herpes simplex virus transcriptional regulatory protein VP16. *Genes Dev,* 13**,** 1692-703.

LOHMANN, J. U., HONG, R. L., HOBE, M., BUSCH, M. A., PARCY, F., SIMON, R. & WEIGEL, D. 2001. A molecular link between stem cell regulation and floral patterning in Arabidopsis. *Cell,* 105**,** 793-803.

LOU, Y., XU, X. F., ZHU, J., GU, J. N., BLACKMORE, S. & YANG, Z. N. 2014. The tapetal AHL family protein TEK determines nexine formation in the pollen wall. *Nat Commun,* 5**,** 3855.

LU, S. F., SUN, Y. H., AMERSON, H. & CHIANG, V. L. 2007. MicroRNAs in loblolly pine (Pinus taeda L.) and their association with fusiform rust gall development. *Plant Journal,* 51**,** 1077-1098.

LUNAU, K. 2004. Adaptive radiation and coevolution - pollination biology case studies. *Organisms Diversity & Evolution,* 4**,** 207-224.

LYONS, E., PEDERSEN, B., KANE, J., ALAM, M., MING, R., TANG, H., WANG, X., BOWERS, J., PATERSON, A., LISCH, D. & FREELING, M. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol,* 148**,** 1772-81.

MA, J. 2011. Reconstructing the History of Large-Scale Genomic Changes: Biological Questions and Computational Challenges. *Journal of Computational Biology,* 18**,** 879-893.

MA, J., PRZIBILLA, E., HU, J., BOGORAD, L. & PTASHNE, M. 1988. Yeast activators stimulate plant gene expression. *Nature,* 334**,** 631-3.

MA, J. & PTASHNE, M. 1987. A new class of yeast transcriptional activators. *Cell,* 51**,** 113-9.

MA, Q., DAI, X., XU, Y., GUO, J., LIU, Y., CHEN, N., XIAO, J., ZHANG, D., XU, Z., ZHANG, X. & CHONG, K. 2009. Enhanced tolerance to chilling stress in OsMYB3R-2 transgenic rice is mediated by alteration in cell cycle and ectopic expression of stress genes. *Plant Physiol,* 150**,** 244-56.

MAGALLON, S., GOMEZ-ACEVEDO, S., SANCHEZ-REYES, L. L. & HERNANDEZ-HERNANDEZ, T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol,* 207**,** 437-53.

MAGALLON, S., HILU, K. W. & QUANDT, D. 2013. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am J Bot,* 100**,** 556-73.

MALKINAPYKH, I. G. 1996. Mathematical model of the growth and development of annual plants based on response functions .1. General description of the model. *Russian Journal of Plant Physiology,* 43**,** 724-732.

MARCHANT, H. J., PICKETT-HEAPS, J. D. & JACOBS, K. 1973. An ultrastructural study of zoosporogenesis and the mature zoospore of Klebsormidium flaccidum. *Cytobios,* 8**,** 95-107.

MARCUSSEN, T., SANDVE, S. R., HEIER, L., SPANNAGL, M., PFEIFER, M., JAKOBSEN, K. S., WULFF, B. B. H., STEUERNAGEL, B., MAYER, K. F. X., OLSEN, O. A. & SEQUENCING, I. W. G. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science,* 345.

MARKMANN-MULISCH, U., HADI, M. Z., KOEPCHEN, K., ALONSO, J. C., RUSSO, V. E. A., SCHELL, J. & REISS, B. 2002. The organization of Physcomitrella patens RAD51 genes is unique among eukaryotic organisms. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 2959-2964.

MARKMANN-MULISCH, U., WENDELER, E., ZOBELL, O., SCHWEEN, G., STEINBISS, H. H. & REISS, B. 2007. Differential requirements for RAD51 in Physcomitrella patens and Arabidopsis thaliana development and DNA damage repair. *Plant Cell,* 19**,** 3080-9.

MARRA, M. A., JONES, S. J., ASTELL, C. R., HOLT, R. A., BROOKS-WILSON, A., BUTTERFIELD, Y. S., KHATTRA, J., ASANO, J. K., BARBER, S. A., CHAN, S. Y., CLOUTIER, A., COUGHLIN, S. M., FREEMAN, D., GIRN, N., GRIFFITH, O. L., LEACH, S. R., MAYO, M., MCDONALD, H., MONTGOMERY, S. B., PANDOH, P. K., PETRESCU, A. S., ROBERTSON, A. G., SCHEIN, J. E., SIDDIQUI, A., SMAILUS, D. E., STOTT, J. M., YANG, G. S., PLUMMER, F., ANDONOV, A., ARTSOB, H., BASTIEN, N., BERNARD, K., BOOTH, T. F., BOWNESS, D., CZUB, M., DREBOT, M., FERNANDO, L., FLICK, R., GARBUTT, M., GRAY, M., GROLLA, A., JONES, S., FELDMANN, H., MEYERS, A., KABANI, A., LI, Y., NORMAND, S., STROHER, U., TIPPLES, G. A., TYLER, S., VOGRIG, R., WARD, D., WATSON, B., BRUNHAM, R. C., KRAJDEN, M., PETRIC, M., SKOWRONSKI, D. M., UPTON, C. & ROPER, R. L. 2003. The Genome sequence of the SARS-associated coronavirus. *Science,* 300**,** 1399-404.

MARTIN, C. & PAZ-ARES, J. 1997. MYB transcription factors in plants. *Trends Genet,* 13**,** 67-73.

MARTINS, M. J., MOTA, C. F. & PEARSON, G. A. 2013. Sex-biased gene expression in the brown alga Fucus vesiculosus. *BMC Genomics,* 14**,** 294.

MASER, P., THOMINE, S., SCHROEDER, J. I., WARD, J. M., HIRSCHI, K., SZE, H., TALKE, I. N., AMTMANN, A., MAATHUIS, F. J. M., SANDERS, D., HARPER, J. F., TCHIEU, J., GRIBSKOV, M., PERSANS, M. W., SALT, D. E., KIM, S. A. & GUERINOT, M. L. 2001. Phylogenetic relationships within cation transporter families of Arabidopsis. *Plant Physiology,* 126**,** 1646-1667.

MASSARI, M. E. & MURRE, C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol,* 20**,** 429-40.

MCCOURT, R. M., DELWICHE, C. F. & KAROL, K. G. 2004. Charophyte algae and land plant origins. *Trends Ecol Evol,* 19**,** 661-6.

MCCOURT, R. M., KAROL, K. G., BELL, J., HELM-BYCHOWSKI, K. M., GRAJEWSKA, A., WOJCIECHOWSKI, M. F. & HOSHAW, R. W. 2000. Phylogeny of the conjugating green algae (Zygnemophyceae) based on rbcL sequences. *Journal of Phycology,* 36**,** 747-758.

MESSIER, W. & STEWART, C. B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature,* 385**,** 151-4.

MEYERS, B. C., AXTELL, M. J., BARTEL, B., BARTEL, D. P., BAULCOMBE, D., BOWMAN, J. L., CAO, X., CARRINGTON, J. C., CHEN, X. M., GREEN, P. J., GRIFFITHS-JONES, S., JACOBSEN, S. E., MALLORY, A. C., MARTIENSSEN, R. A., POETHIG, R. S., QI, Y. J., VAUCHERET, H., VOINNET, O., WATANABE, Y., WEIGEL, D. & ZHUI, J. K. 2008. Criteria for Annotation of Plant MicroRNAs. *Plant Cell,* 20**,** 3186-3190.

MILBRADT, A. G., KULKARNI, M., YI, T., TAKEUCHI, K., SUN, Z. Y., LUNA, R. E., SELENKO, P., NAAR, A. M. & WAGNER, G. 2011. Structure of the VP16 transactivator target in the Mediator. *Nat Struct Mol Biol,* 18**,** 410-5.

MILLAR, A. A. & GUBLER, F. 2005. The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. *Plant Cell,* 17**,** 705-21.

MISSAOUI, A. M., PATERSON, A. H. & BOUTON, J. H. 2005. Investigation of genomic organization in switchgrass (Panicum virgatum L.) using DNA markers. *Theor Appl Genet,* 110**,** 1372-83.

MITCHELL, P. J. & TJIAN, R. 1989. Transcriptional Regulation in Mammalian-Cells by Sequence-Specific DNA-Binding Proteins. *Science,* 245**,** 371-378.

MOLNAR, A., SCHWACH, F., STUDHOLME, D. J., THUENEMANN, E. C. & BAULCOMBE, D. C. 2007. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature,* 447**,** 1126-9.

MOMEN-ROKNABADI, A., SADEGHI, M., PEZESHK, H. & MARASHI, S. A. 2008. Impact of residue accessible surface area on the prediction of protein secondary structures. *BMC Bioinformatics,* 9**,** 357.

MOORE, M. J., BELL, C. D., SOLTIS, P. S. & SOLTIS, D. E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci U S A,* 104**,** 19363-8.

MOORE, M. J., SOLTIS, P. S., BELL, C. D., BURLEIGH, J. G. & SOLTIS, D. E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A,* 107**,** 4623-8.

MORI, T., IGAWA, T., TAMIYA, G., MIYAGISHIMA, S. Y. & BERGER, F. 2014. Gamete attachment requires GEX2 for successful fertilization in Arabidopsis. *Curr Biol,* 24**,** 170-5.

MORRAN, L. T., SCHMIDT, O. G., GELARDEN, I. A., PARRISH, R. C. & LIVELY, C. M. 2011. Running with the Red Queen: Host-Parasite Coevolution Selects for Biparental Sex. *Science,* 333**,** 216-218.

MUHLHAUSEN, S. & KOLLMAR, M. 2013. Whole genome duplication events in plant evolution reconstructed and predicted using myosin motor proteins. *Bmc Evolutionary Biology,* 13.

MÜLLER, S. J., GÜTLE, D. D., JACQUOT, J.-P. & RESKI, R. 2015. Can mosses serve as model organisms for forest research? *Annals of Forest Science*.

NAKATA, Y., SHETZLINE, S., SAKASHITA, C., KALOTA, A., RALLAPALLI, R., RUDNICK, S. I., ZHANG, Y., EMERSON, S. G. & GEWIRTZ, A. M. 2007. c-Myb contributes to G2/M cell cycle transition in human hematopoietic cells by direct regulation of cyclin B1 expression. *Mol Cell Biol,* 27**,** 2048-58.

NEI, M. & KUMAR, S. 2000. *Molecular evolution and phylogenetics,* New York, Oxford University Press.

NELSON, M. R., BAND, L. R., DYSON, R. J., LESSINNES, T., WELLS, D. M., YANG, C., EVERITT, N. M., JENSEN, O. E. & WILSON, Z. A. 2012. A biomechanical model of anther opening reveals the roles of dehydration and secondary thickening. *New Phytol,* 196**,** 1030-7.

NEWMAN, L. J., PERAZZA, D. E., JUDA, L. & CAMPBELL, M. M. 2004. Involvement of the R2R3-MYB, AtMYB61, in the ectopic lignification and dark-photomorphogenic components of the det3 mutant phenotype. *Plant J,* 37**,** 239-50.

NICKRENT, D. L., PARKINSON, C. L., PALMER, J. D. & DUFF, R. J. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol Biol Evol,* 17**,** 1885-95.

NIELSEN, R. & YANG, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics,* 148**,** 929-36.

NIKOLOV, D. B. & BURLEY, S. K. 1997. RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A,* 94**,** 15-22.

NISHIYAMA, T., FUJITA, T., SHIN, I. T., SEKI, M., NISHIDE, H., UCHIYAMA, I., KAMIYA, A., CARNINCI, P., HAYASHIZAKI, Y., SHINOZAKI, K., KOHARA, Y. & HASEBE, M. 2003. Comparative

genomics of Physcomitrella patens gametophytic transcriptome and Arabidopsis thaliana: implication for land plant evolution. *Proc Natl Acad Sci U S A,* 100**,** 8007-12.

NISHIYAMA, T., WOLF, P. G., KUGITA, M., SINCLAIR, R. B., SUGITA, M., SUGIURA, C., WAKASUGI, T., YAMADA, K., YOSHINAGA, K., YAMAGUCHI, K., UEDA, K. & HASEBE, M. 2004. Chloroplast phylogeny indicates that bryophytes are monophyletic. *Mol Biol Evol,* 21**,** 1813-9.

NYSTEDT, B., STREET, N. R., WETTERBOM, A., ZUCCOLO, A., LIN, Y. C., SCOFIELD, D. G., VEZZI, F., DELHOMME, N., GIACOMELLO, S., ALEXEYENKO, A., VICEDOMINI, R., SAHLIN, K., SHERWOOD, E., ELFSTRAND, M., GRAMZOW, L., HOLMBERG, K., HALLMAN, J., KEECH, O., KLASSON, L., KORIABINE, M., KUCUKOGLU, M., KALLER, M., LUTHMAN, J., LYSHOLM, F., NIITTYLA, T., OLSON, A., RILAKOVIC, N., RITLAND, C., ROSSELLO, J. A., SENA, J., SVENSSON, T., TALAVERA-LOPEZ, C., THEISSEN, G., TUOMINEN, H., VANNESTE, K., WU, Z. Q., ZHANG, B., ZERBE, P., ARVESTAD, L., BHALERAO, R., BOHLMANN, J., BOUSQUET, J., GARCIA GIL, R., HVIDSTEN, T. R., DE JONG, P., MACKAY, J., MORGANTE, M., RITLAND, K., SUNDBERG, B., THOMPSON, S. L., VAN DE PEER, Y., ANDERSSON, B., NILSSON, O., INGVARSSON, P. K., LUNDEBERG, J. & JANSSON, S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature,* 497**,** 579-84.

ODA, K., YAMATO, K., OHTA, E., NAKAMURA, Y., TAKEMURA, M., NOZATO, N., AKASHI, K., KANEGAE, T., OGURA, Y., KOHCHI, T. & ET AL. 1992. Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA. A primitive form of plant mitochondrial genome. *J Mol Biol,* 223**,** 1-7.

OH, I. H. & REDDY, E. P. 1999. The myb gene family in cell growth, differentiation and apoptosis. *Oncogene,* 18**,** 3017-33.

OHASHI-ITO, K. & BERGMANN, D. C. 2006. Arabidopsis FAMA controls the final proliferation/differentiation switch during stomatal development. *Plant Cell,* 18**,** 2493-505.

OHI, R., FEOKTISTOVA, A., MCCANN, S., VALENTINE, V., LOOK, A. T., LIPSICK, J. S. & GOULD, K. L. 1998. Myb-related Schizosaccharomyces pombe cdc5p is structurally and functionally conserved in eukaryotes (vol 18, pg 4097, 1998). *Molecular and Cellular Biology,* 18**,** 6164-6164.

OHI, R., MCCOLLUM, D., HIRANI, B., DEN HAESE, G. J., ZHANG, X., BURKE, J. D., TURNER, K. & GOULD, K. L. 1994. The Schizosaccharomyces pombe cdc5+ gene encodes an essential protein with homology to c-Myb. *EMBO J,* 13**,** 471-83.

OHTA, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature,* 246**,** 96-8.

OHTA, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics,* 23**,** 263-286.

OHTA, T. 2011. Near-neutrality, robustness, and epigenetics. *Genome Biol Evol,* 3**,** 1034-8.

OHYAMA, K., FUKUZAWA, H., KOHCHI, T., SHIRAI, H., SANO, T., SANO, S., UMESONO, K., SHIKI, Y., TAKEUCHI, M., CHANG, Z., AOTA, S., INOKUCHI, H. & OZEKI, H. 1986. Chloroplast Gene Organization Deduced from Complete Sequence of Liverwort Marchantia-Polymorpha Chloroplast DNA. *Nature,* 322**,** 572-574.

OKADA, T., BHALLA, P. L. & SINGH, M. B. 2006. Expressed sequence tag analysis of Lilium longiflorum generative cells. *Plant Cell Physiol,* 47**,** 698-705.

OKADA, T., ENDO, M., SINGH, M. B. & BHALLA, P. L. 2005. Analysis of the histone H3 gene family in Arabidopsis and identification of the male-gamete-specific variant AtMGH3. *Plant J,* 44**,** 557-68.

OPPENHEIMER, D. G., HERMAN, P. L., SIVAKUMARAN, S., ESCH, J. & MARKS, M. D. 1991. A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. *Cell,* 67**,** 483-93.

ORTIZ-RAMIREZ, C., HERNANDEZ-CORONADO, M., THAMM, A., CATARINO, B., WANG, M., DOLAN, L., FEIJO, J. A. & BECKER, J. D. 2016. A Transcriptome Atlas of Physcomitrella

patens Provides Insights into the Evolution and Development of Land Plants. *Mol Plant,* 9**,** 205-20.

PĂCURAR, D. I., THORDAL-CHRISTENSEN, H., PĂCURAR, M. L., PAMFIL, D., BOTEZ, C. & BELLINI, C. 2011. Agrobacterium tumefaciens: From crown gall tumors to genetic transformation. *Physiological and Molecular Plant Pathology,* 76**,** 76-81.

PALATNIK, J. F., ALLEN, E., WU, X. L., SCHOMMER, C., SCHWAB, R., CARRINGTON, J. C. & WEIGEL, D. 2003. Control of leaf morphogenesis by microRNAs. *Nature,* 425**,** 257-263.

PALATNIK, J. F., WOLLMANN, H., SCHOMMER, C., SCHWAB, R., BOISBOUVIER, J., RODRIGUEZ, R., WARTHMANN, N., ALLEN, E., DEZULIAN, T., HUSON, D., CARRINGTON, J. C. & WEIGEL, D. 2007. Sequence and expression differences underlie functional specialization of Arabidopsis microRNAs miR159 and miR319. *Dev Cell,* 13**,** 115-25.

PALMA, K., ZHAO, Q., CHENG, Y. T., BI, D., MONAGHAN, J., CHENG, W., ZHANG, Y. & LI, X. 2007. Regulation of plant innate immunity by three proteins in a complex conserved across the plant and animal kingdoms. *Genes Dev,* 21**,** 1484-93.

PARK, S. K., HOWDEN, R. & TWELL, D. 1998. The Arabidopsis thaliana gametophytic mutation gemini pollen1 disrupts microspore polarity, division asymmetry and pollen cell fate. *Development,* 125**,** 3789-3799.

PARKER, J. D. 2004. A major evolutionary transition to more than two sexes? *Trends in Ecology & Evolution,* 19**,** 83-86.

PARSCH, J. & ELLEGREN, H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat Rev Genet,* 14**,** 83-7.

PATEN, B., HERRERO, J., FITZGERALD, S., BEAL, K., FLICEK, P., HOLMES, I. & BIRNEY, E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res,* 18**,** 1829-43.

PATERSON, A. H., BOWERS, J. E., BRUGGMANN, R., DUBCHAK, I., GRIMWOOD, J., GUNDLACH, H., HABERER, G., HELLSTEN, U., MITROS, T., POLIAKOV, A., SCHMUTZ, J., SPANNAGL, M., TANG, H., WANG, X., WICKER, T., BHARTI, A. K., CHAPMAN, J., FELTUS, F. A., GOWIK, U., GRIGORIEV, I. V., LYONS, E., MAHER, C. A., MARTIS, M., NARECHANIA, A., OTILLAR, R. P., PENNING, B. W., SALAMOV, A. A., WANG, Y., ZHANG, L., CARPITA, N. C., FREELING, M., GINGLE, A. R., HASH, C. T., KELLER, B., KLEIN, P., KRESOVICH, S., MCCANN, M. C., MING, R., PETERSON, D. G., MEHBOOB UR, R., WARE, D., WESTHOFF, P., MAYER, K. F., MESSING, J. & ROKHSAR, D. S. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature,* 457**,** 551-6.

PAUKLIN, S., MADRIGAL, P., BERTERO, A. & VALLIER, L. 2016. Initiation of stem cell differentiation involves cell cycle-dependent regulation of developmental genes by Cyclin D. *Genes Dev,* 30**,** 421-33.

PAYNE, C. T., ZHANG, F. & LLOYD, A. M. 2000. GL3 encodes a bHLH protein that regulates trichome development in arabidopsis through interaction with GL1 and TTG1. *Genetics,* 156**,** 1349-1362.

PAYNE, T., CLEMENT, J., ARNOLD, D. & LLOYD, A. 1999. Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in Nicotiana tabacum. *Development,* 126**,** 671-82.

PAZ-ARES, J., GHOSAL, D., WIENAND, U., PETERSON, P. A. & SAEDLER, H. 1987. The regulatory c1 locus of Zea mays encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J,* 6**,** 3553-8.

PEARCE, S., FERGUSON, A., KING, J. & WILSON, Z. A. 2015. FlowerNet: a gene expression correlation network for anther and pollen development. *Plant Physiol,* 167**,** 1717-30.

PEDERSEN, J. S., BEJERANO, G., SIEPEL, A., ROSENBLOOM, K., LINDBLAD-TOH, K., LANDER, E. S., KENT, J., MILLER, W. & HAUSSLER, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol,* 2**,** e33.

PENFIELD, S., MEISSNER, R. C., SHOUE, D. A., CARPITA, N. C. & BEVAN, M. W. 2001. MYB61 is required for mucilage deposition and extrusion in the Arabidopsis seed coat. *Plant Cell,* 13**,** 2777-2791.

PETERS, B., CASEY, J., AIDLEY, J., ZOHRAB, S., BORG, M., TWELL, D. & BROWNFIELD, L. 2016. A cis-regulatory module in the transcription factor DUO1 promoter. *Plant Physiol*.

PETERS, C. W., SIPPEL, A. E., VINGRON, M. & KLEMPNAUER, K. H. 1987. Drosophila and vertebrate myb proteins share two conserved regions, one of which functions as a DNA-binding domain. *EMBO J,* 6**,** 3085-90.

PETRONI, K., FALASCA, G., CALVENZANI, V., ALLEGRA, D., STOLFI, C., FABRIZI, L., ALTAMURA, M. M. & TONELLI, C. 2008. The AtMYB11 gene from Arabidopsis is expressed in meristematic cells and modulates growth in planta and organogenesis in vitro. *J Exp Bot,* 59**,** 1201-13.

PHAN, H. A., IACUONE, S., LI, S. F. & PARISH, R. W. 2011. The MYB80 Transcription Factor Is Required for Pollen Development and the Regulation of Tapetal Programmed Cell Death in Arabidopsis thaliana. *Plant Cell,* 23**,** 2209-2224.

PHAN, H. A., LI, S. F. & PARISH, R. W. 2012. MYB80, a regulator of tapetal and pollen development, is functionally conserved in crops. *Plant Molecular Biology,* 78**,** 171-183.

PICKETTHEAPS, J. D. & WETHERBEE, R. 1987. Spindle Function in the Green-Alga Mougeotia - Absence of Anaphase a Correlates with Postmitotic Nuclear Migration. *Cell Motility and the Cytoskeleton,* 7**,** 68-77.

PISKACEK, M., HAVELKA, M., REZACOVA, M. & KNIGHT, A. 2016. The 9aaTAD Transactivation Domains: From Gal4 to p53. *PLoS One,* 11**,** e0162842.

PISKACEK, M., VASKU, A., HAJEK, R. & KNIGHT, A. 2015. Shared structural features of the 9aaTAD family in complex with CBP. *Mol Biosyst,* 11**,** 844-51.

PISKACEK, S., GREGOR, M., NEMETHOVA, M., GRABNER, M., KOVARIK, P. & PISKACEK, M. 2007. Nine-amino-acid transactivation domain: establishment and prediction utilities. *Genomics,* 89**,** 756-68.

PLACKETT, A. R., THOMAS, S. G., WILSON, Z. A. & HEDDEN, P. 2011. Gibberellin control of stamen development: a fertile field. *Trends Plant Sci,* 16**,** 568-78.

PLACKETT, A. R. G., POWERS, S. J., FERNANDEZ-GARCIA, N., URBANOVA, T., TAKEBAYASHI, Y., SEO, M., JIKUMARU, Y., BENLLOCH, R., NILSSON, O., RUIZ-RIVERO, O., PHILLIPS, A. L., WILSON, Z. A., THOMAS, S. G. & HEDDEN, P. 2012. Analysis of the Developmental Roles of the Arabidopsis Gibberellin 20-Oxidases Demonstrates That GA20ox1, -2, and -3 Are the Dominant Paralogs. *Plant Cell,* 24**,** 941-960.

POTATO GENOME SEQUENCING, C., XU, X., PAN, S., CHENG, S., ZHANG, B., MU, D., NI, P., ZHANG, G., YANG, S., LI, R., WANG, J., ORJEDA, G., GUZMAN, F., TORRES, M., LOZANO, R., PONCE, O., MARTINEZ, D., DE LA CRUZ, G., CHAKRABARTI, S. K., PATIL, V. U., SKRYABIN, K. G., KUZNETSOV, B. B., RAVIN, N. V., KOLGANOVA, T. V., BELETSKY, A. V., MARDANOV, A. V., DI GENOVA, A., BOLSER, D. M., MARTIN, D. M., LI, G., YANG, Y., KUANG, H., HU, Q., XIONG, X., BISHOP, G. J., SAGREDO, B., MEJIA, N., ZAGORSKI, W., GROMADKA, R., GAWOR, J., SZCZESNY, P., HUANG, S., ZHANG, Z., LIANG, C., HE, J., LI, Y., HE, Y., XU, J., ZHANG, Y., XIE, B., DU, Y., QU, D., BONIERBALE, M., GHISLAIN, M., HERRERA MDEL, R., GIULIANO, G., PIETRELLA, M., PERROTTA, G., FACELLA, P., O'BRIEN, K., FEINGOLD, S. E., BARREIRO, L. E., MASSA, G. A., DIAMBRA, L., WHITTY, B. R., VAILLANCOURT, B., LIN, H., MASSA, A. N., GEOFFROY, M., LUNDBACK, S., DELLAPENNA, D., BUELL, C. R., SHARMA, S. K., MARSHALL, D. F., WAUGH, R., BRYAN, G. J., DESTEFANIS, M., NAGY, I., MILBOURNE, D., THOMSON, S. J., FIERS, M., JACOBS, J. M., NIELSEN, K. L., SONDERKAER, M., IOVENE, M., TORRES, G. A., JIANG, J., VEILLEUX, R. E., BACHEM, C. W., DE BOER, J., BORM, T., KLOOSTERMAN, B., VAN ECK, H., DATEMA, E., HEKKERT, B., GOVERSE, A., VAN HAM, R. C. & VISSER, R. G. 2011. Genome sequence and analysis of the tuber crop potato. *Nature,* 475**,** 189-95.

PTASHNE, M. & GANN, A. 1997. Transcriptional activation by recruitment. *Nature,* 386**,** 569-77.

PUCCI, B., KASTEN, M. & GIORDANO, A. 2000. Cell cycle and apoptosis. *Neoplasia,* 2**,** 291-9.

QIU, Y.-L., LI, L., WANG, B., XUE, J.-Y., HENDRY, T. A., LI, R.-Q., BROWN, J. W., LIU, Y., HUDSON, G. T. & CHEN, Z.-D. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution,* 48**,** 391-425.

QIU, Y. L., CHO, Y. R., COX, J. C. & PALMER, J. D. 1998. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature,* 394**,** 671-674.

QIU, Y. L. & LEE, J. 2000. Transition to a land flora: A molecular phylogenetic perspective. *Journal of Phycology,* 36**,** 799-802.

QIU, Y. L., LI, L., WANG, B., CHEN, Z., KNOOP, V., GROTH-MALONEK, M., DOMBROVSKA, O., LEE, J., KENT, L., REST, J., ESTABROOK, G. F., HENDRY, T. A., TAYLOR, D. W., TESTA, C. M., AMBROS, M., CRANDALL-STOTLER, B., DUFF, R. J., STECH, M., FREY, W., QUANDT, D. & DAVIS, C. C. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. *Proc Natl Acad Sci U S A,* 103**,** 15511-6.

QIU, Y. L., LI, L. B., WANG, B., CHEN, Z. D., DOMBROVSKA, O., LEE, J., KENT, L., LI, R. Q., JOBSON, R. W., HENDRY, T. A., TAYLOR, D. W., TESTA, C. M. & AMBROS, M. 2007. A nonflowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial, and nuclear genes. *International Journal of Plant Sciences,* 168**,** 691-708.

RABINOWICZ, P. D., BRAUN, E. L., WOLFE, A. D., BOWEN, B. & GROTEWOLD, E. 1999. Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. *Genetics,* 153**,** 427-44.

RAJEWSKY, N. 2006. microRNA target predictions in animals. *Nat Genet,* 38 Suppl**,** S8-13.

RAMBAUT, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics,* 16**,** 395-399.

RAMBO, R. P., WILLIAMS, G. J. & TAINER, J. A. 2010. Achieving fidelity in homologous recombination despite extreme complexity: informed decisions by molecular profiling. *Mol Cell,* 40**,** 347-8.

RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc,* 8**,** 2281-308.

RANJAN, A., DICKOPF, S., ULLRICH, K. K., RENSING, S. A. & HOECKER, U. 2014. Functional analysis of COP1 and SPA orthologs from Physcomitrella and rice during photomorphogenesis of transgenic Arabidopsis reveals distinct evolutionary conservation. *BMC Plant Biol,* 14**,** 178.

REGIER, J. L., SHEN, F. & TRIEZENBERG, S. J. 1993. Pattern of aromatic and hydrophobic amino acids critical for one of two subdomains of the VP16 transcriptional activator. *Proc Natl Acad Sci U S A,* 90**,** 883-7.

RENAK, D., DUPL'AKOVA, N. & HONYS, D. 2012. Wide-scale screening of T-DNA lines for transcription factor genes affecting male gametophyte development in Arabidopsis. *Sex Plant Reprod,* 25**,** 39-60.

RENNE, P. R., DEINO, A. L., HILGEN, F. J., KUIPER, K. F., MARK, D. F., MITCHELL, W. S., 3RD, MORGAN, L. E., MUNDIL, R. & SMIT, J. 2013. Time scales of critical events around the Cretaceous-Paleogene boundary. *Science,* 339**,** 684-7.

RENSING, S. A., LANG, D., ZIMMER, A. D., TERRY, A., SALAMOV, A., SHAPIRO, H., NISHIYAMA, T., PERROUD, P. F., LINDQUIST, E. A., KAMISUGI, Y., TANAHASHI, T., SAKAKIBARA, K., FUJITA, T., OISHI, K., SHIN, I. T., KUROKI, Y., TOYODA, A., SUZUKI, Y., HASHIMOTO, S., YAMAGUCHI, K., SUGANO, S., KOHARA, Y., FUJIYAMA, A., ANTEROLA, A., AOKI, S., ASHTON, N., BARBAZUK, W. B., BARKER, E., BENNETZEN, J. L., BLANKENSHIP, R., CHO, S. H., DUTCHER, S. K., ESTELLE, M., FAWCETT, J. A., GUNDLACH, H., HANADA, K., HEYL, A., HICKS, K. A., HUGHES, J., LOHR, M., MAYER, K., MELKOZERNOV, A., MURATA, T., NELSON, D. R., PILS, B., PRIGGE, M., REISS, B., RENNER, T., ROMBAUTS, S., RUSHTON, P.

J., SANDERFOOT, A., SCHWEEN, G., SHIU, S. H., STUEBER, K., THEODOULOU, F. L., TU, H., VAN DE PEER, Y., VERRIER, P. J., WATERS, E., WOOD, A., YANG, L., COVE, D., CUMING, A. C., HASEBE, M., LUCAS, S., MISHLER, B. D., RESKI, R., GRIGORIEV, I. V., QUATRANO, R. S. & BOORE, J. L. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science,* 319**,** 64-9.

REYES, J. L. & CHUA, N. H. 2007. ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. *Plant J,* 49**,** 592-606.

RHOADES, M. W., REINHART, B. J., LIM, L. P., BURGE, C. B., BARTEL, B. & BARTEL, D. P. 2002. Prediction of plant microRNA targets. *Cell,* 110**,** 513-20.

RIECHMANN, J. L. 2000. Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science,* 290**,** 2105-2110.

RINDI, F., GUIRY, M. D. & LÓPEZ-BAUTISTA, J. M. 2008. Distribution, Morphology, and Phylogeny Ofklebsormidium(Klebsormidiales, Charophyceae) in Urban Environments in Europe1. *Journal of Phycology,* 44**,** 1529-1540.

ROEDER, R. G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci,* 21**,** 327-35.

ROGERS, C. E., MATTOX, K. R. & STEWART, K. D. 1980. The Zoospore of Chlorokybus-Atmophyticus, a Charophyte with Sarcinoid Growth Habit. *American Journal of Botany,* 67**,** 774-783.

ROMERO, I., FUERTES, A., BENITO, M. J., MALPICA, J. M., LEYVA, A. & PAZ-ARES, J. 1998. More than 80R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. *Plant Journal,* 14**,** 273-284.

ROTMAN, N., DURBARRY, A., WARDLE, A., YANG, W. C., CHABOUD, A., FAURE, J. E., BERGER, F. & TWELL, D. 2005. A novel class of MYB factors controls sperm-cell formation in plants. *Curr Biol,* 15**,** 244-8.

RUBINSTEIN, C. V., GERRIENNE, P., DE LA PUENTE, G. S., ASTINI, R. A. & STEEMANS, P. 2010. Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol,* 188**,** 365-9.

RUBY, J. G., JAN, C., PLAYER, C., AXTELL, M. J., LEE, W., NUSBAUM, C., GE, H. & BARTEL, D. P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell,* 127**,** 1193-207.

RUBY, J. G., STARK, A., JOHNSTON, W. K., KELLIS, M., BARTEL, D. P. & LAI, E. C. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res,* 17**,** 1850-64.

RUHFEL, B. R., GITZENDANNER, M. A., SOLTIS, P. S., SOLTIS, D. E. & BURLEIGH, J. G. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol Biol,* 14**,** 23.

SADOWSKI, I., MA, J., TRIEZENBERG, S. & PTASHNE, M. 1988. Gal4-Vp16 Is an Unusually Potent Transcriptional Activator. *Nature,* 335**,** 563-564.

SAIKUMAR, P., MURALI, R. & REDDY, E. P. 1990. Role of tryptophan repeats and flanking amino acids in Myb-DNA interactions. *Proc Natl Acad Sci U S A,* 87**,** 8452-6.

SAINZ, M. B., GOFF, S. A. & CHANDLER, V. L. 1997. Extensive mutagenesis of a transcriptional activation domain identifies single hydrophobic and acidic amino acids important for activation in vivo. *Mol Cell Biol,* 17**,** 115-22.

SALIPANTE, S. J. & HORWITZ, M. S. 2006. Phylogenetic fate mapping. *Proc Natl Acad Sci U S A,* 103**,** 5448-53.

SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution,* 14**,** 1218-1231.

SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution,* 19**,** 101-109.

SANDERSON, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics,* 19**,** 301-2.

SANDHOLZER, J., HOETH, M., PISKACEK, M., MAYER, H. & DE MARTIN, R. 2007. A novel 9-amino-acid transactivation domain in the C-terminal part of Sox18. *Biochem Biophys Res Commun,* 360**,** 370-4.

SARI, U. 2015. *Pollen polymorphism in Magnoliaceae and conservation of plant male germline regulators.* Thesis (Ph.D.), University of Leicester.

SAUQUET, H., VON BALTHAZAR, M., MAGALLON, S., DOYLE, J. A., ENDRESS, P. K., BAILES, E. J., BARROSO DE MORAIS, E., BULL-HERENU, K., CARRIVE, L., CHARTIER, M., CHOMICKI, G., COIRO, M., CORNETTE, R., EL OTTRA, J. H. L., EPICOCO, C., FOSTER, C. S. P., JABBOUR, F., HAEVERMANS, A., HAEVERMANS, T., HERNANDEZ, R., LITTLE, S. A., LOFSTRAND, S., LUNA, J. A., MASSONI, J., NADOT, S., PAMPERL, S., PRIEU, C., REYES, E., DOS SANTOS, P., SCHOONDERWOERD, K. M., SONTAG, S., SOULEBEAU, A., STAEDLER, Y., TSCHAN, G. F., WING-SZE LEUNG, A. & SCHONENBERGER, J. 2017. The ancestral flower of angiosperms and its early diversification. *Nat Commun,* 8**,** 16047.

SAVIR, Y. & TLUSTY, T. 2007. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PLoS One,* 2**,** e468.

SAVIR, Y. & TLUSTY, T. 2010. RecA-Mediated Homology Search as a Nearly Optimal Signal Detection System. *Molecular Cell,* 40**,** 388-396.

SAVIR, Y. & TLUSTY, T. 2013. The Ribosome as an Optimal Decoder: A Lesson in Molecular Recognition. *Cell,* 153**,** 471-479.

SCHAEFER, D., ZRYD, J. P., KNIGHT, C. D. & COVE, D. J. 1991. Stable Transformation of the Moss Physcomitrella-Patens. *Molecular & General Genetics,* 226**,** 418-424.

SCHAEFER, D. G. 2002. A new moss genetics: targeted mutagenesis in Physcomitrella patens. *Annu Rev Plant Biol,* 53**,** 477-501.

SCHAEFER, D. G., DELACOTE, F., CHARLOT, F., VRIELYNCK, N., GUYON-DEBAST, A., LE GUIN, S., NEUHAUS, J. M., DOUTRIAUX, M. P. & NOGUE, F. 2010. RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens. *DNA Repair (Amst),* 9**,** 526-33.

SCHAEFER, D. G. & ZRYD, J. P. 1997. Efficient gene targeting in the moss Physcomitrella patens. *Plant Journal,* 11**,** 1195-1206.

SCHIML, S., FAUSER, F. & PUCHTA, H. 2014. The CRISPR/Cas system can be used as nuclease for in planta gene targeting and as paired nickases for directed mutagenesis in Arabidopsis resulting in heritable progeny. *Plant J,* 80**,** 1139-50.

SCHMITZ, G., TILLMANN, E., CARRIERO, F., FIORE, C., CELLINI, F. & THERES, K. 2002. The tomato Blind gene encodes a MYB transcription factor that controls the formation of lateral meristems. *Proc Natl Acad Sci U S A,* 99**,** 1064-9.

SCHNEIDER, H., SCHUETTPELZ, E., PRYER, K. M., CRANFILL, R., MAGALLON, S. & LUPIA, R. 2004. Ferns diversified in the shadow of angiosperms. *Nature,* 428**,** 553-7.

SCHWEDE, T., KOPP, J., GUEX, N. & PEITSCH, M. C. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research,* 31**,** 3381-3385.

SEITZ, H. 2009. Redefining microRNA targets. *Curr Biol,* 19**,** 870-3.

SEO, H. S., YANG, J. Y., ISHIKAWA, M., BOLLE, C., BALLESTEROS, M. L. & CHUA, N. H. 2003. LAF1 ubiquitination by COP1 controls photomorphogenesis and is stimulated by SPA1. *Nature,* 423**,** 995-9.

SERJEANT, G. R. 2010. One hundred years of sickle cell disease. *Br J Haematol,* 151**,** 425-9.

SHAO, Z., YIN, J., CHAPMAN, K., GRZEMSKA, M., CLARK, L., WANG, J. & ROSENBAUM, D. M. 2016. High-resolution crystal structure of the human CB1 cannabinoid receptor. *Nature*.

SHARP, P. A. & BURGE, C. B. 1997. Classification of introns: U2-type or U12-type. *Cell,* 91**,** 875-9.

SHAW, A. J., SZOVENYI, P. & SHAW, B. 2011. Bryophyte diversity and evolution: windows into the early evolution of land plants. *Am J Bot,* 98**,** 352-69.

SILVEIRA, A. B., GAUER, L., TOMAZ, J. P., CARDOSO, P. R., CARMELLO-GUERREIRO, S. & VINCENTZ, M. 2007. The Arabidopsis AtbZIP9 protein fused to the VP16 transcriptional activation domain alters leaf and vascular development. *Plant Science,* 172**,** 1148-1156.

SKRIVER, K., OLSEN, F. L., ROGERS, J. C. & MUNDY, J. 1991. cis-acting DNA elements responsive to gibberellin and its antagonist abscisic acid. *Proc Natl Acad Sci U S A,* 88**,** 7266-70.

SLOWINSKI, J. B. & PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology,* 48**,** 814-825.

SMIDKOVA, M., HOLA, M. & ANGELIS, K. J. 2010. Efficient biolistic transformation of the moss Physcomitrella patens. *Biologia Plantarum,* 54**,** 777-780.

SOLANO, R., FUERTES, A., SANCHEZ-PULIDO, L., VALENCIA, A. & PAZ-ARES, J. 1997. A single residue substitution causes a switch from the dual DNA binding specificity of plant transcription factor MYB.Ph3 to the animal c-MYB specificity. *J Biol Chem,* 272**,** 2889-95.

SOLANO, R., NIETO, C., AVILA, J., CANAS, L., DIAZ, I. & PAZ-ARES, J. 1995. Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrida. *EMBO J,* 14**,** 1773-84.

SOLTIS, D. E., ALBERT, V. A., LEEBENS-MACK, J., BELL, C. D., PATERSON, A. H., ZHENG, C., SANKOFF, D., DEPAMPHILIS, C. W., WALL, P. K. & SOLTIS, P. S. 2009. Polyploidy and angiosperm diversification. *Am J Bot,* 96**,** 336-48.

SOLTIS, D. E., BELL, C. D., KIM, S. & SOLTIS, P. S. 2008. Origin and early evolution of angiosperms. *Ann N Y Acad Sci,* 1133**,** 3-25.

SOLTIS, D. E., SMITH, S. A., CELLINESE, N., WURDACK, K. J., TANK, D. C., BROCKINGTON, S. F., REFULIO-RODRIGUEZ, N. F., WALKER, J. B., MOORE, M. J., CARLSWARD, B. S., BELL, C. D., LATVIS, M., CRAWLEY, S., BLACK, C., DIOUF, D., XI, Z., RUSHWORTH, C. A., GITZENDANNER, M. A., SYTSMA, K. J., QIU, Y. L., HILU, K. W., DAVIS, C. C., SANDERSON, M. J., BEAMAN, R. S., OLMSTEAD, R. G., JUDD, W. S., DONOGHUE, M. J. & SOLTIS, P. S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot,* 98**,** 704-30.

SPARKES, I. A., RUNIONS, J., KEARNS, A. & HAWES, C. 2006. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat Protoc,* 1**,** 2019-25.

STARK, A., BRENNECKE, J., RUSSELL, R. B. & COHEN, S. M. 2003. Identification of Drosophila MicroRNA targets. *PLoS Biol,* 1**,** E60.

STAUFFER, D., DE OLIVEIRA, P. M., DE OLIVEIRA, S. M., PENNA, T. J. & MARTINS, J. S. 2001. Computer simulations for biological aging and sexual reproduction. *An Acad Bras Cienc,* 73**,** 15-32.

STEARNS, F. W. & FENSTER, C. B. 2013. Evidence for parallel adaptation to climate across the natural range of Arabidopsis thaliana. *Ecol Evol,* 3**,** 2241-50.

STEEMANS, P., HERISSE, A. L., MELVIN, J., MILLER, M. A., PARIS, F., VERNIERS, J. & WELLMAN, C. H. 2009. Origin and radiation of the earliest vascular land plants. *Science,* 324**,** 353.

STEINER-LANGE, S., UNTE, U. S., ECKSTEIN, L., YANG, C., WILSON, Z. A., SCHMELZER, E., DEKKER, K. & SAEDLER, H. 2003. Disruption of Arabidopsis thaliana MYB26 results in male sterility due to non-dehiscent anthers. *Plant J,* 34**,** 519-28.

STORGAARD, M., DIDION, T., OKKELS, F. & NIELSE, K. K. 2002. Expression of a 434:VP16 chimeric activator leads to high-level activation of gene expression in stable transformants of Arabidopsis. *Transgenic Res,* 11**,** 151-9.

STRACKE, R., WERBER, M. & WEISSHAAR, B. 2001. The R2R3-MYB gene family in Arabidopsis thaliana. *Curr Opin Plant Biol,* 4**,** 447-56.

STROME, S. & LEHMANN, R. 2007. Germ versus soma decisions: lessons from flies and worms. *Science,* 316**,** 392-3.

SUGIMOTO, K., TAKEDA, S. & HIROCHIKA, H. 2000. MYB-related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the tobacco retrotransposon Tto1 and defense-related genes. *Plant Cell,* 12**,** 2511-2528.

SULLIVAN, S. M., HORN, P. J., OLSON, V. A., KOOP, A. H., NIU, W., EBRIGHT, R. H. & TRIEZENBERG, S. J. 1998. Mutational analysis of a transcriptional activation region of the VP16 protein of herpes simplex virus. *Nucleic Acids Res,* 26**,** 4487-96.

SUNG, P. & ROBBERSON, D. L. 1995. DNA Strand Exchange Mediated by a Rad51-Ssdna Nucleoprotein Filament with Polarity Opposite to That of Reca. *Cell,* 82**,** 453-461.

SUREK, B., BEEMELMANNS, U., MELKONIAN, M. & BHATTACHARYA, D. 1994. Ribosomal-Rna Sequence Comparisons Demonstrate an Evolutionary Relationship between Zygnematales and Charophytes. *Plant Systematics and Evolution,* 191**,** 171-181.

SUZUKI, Y. & GOJOBORI, T. 1999. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution,* 16**,** 1315-1328.

SWANSON, W. J. & VACQUIER, V. D. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet,* 3**,** 137-44.

TAHIROV, T. H., SATO, K., ICHIKAWA-IWATA, E., SASAKI, M., INOUE-BUNGO, T., SHIINA, M., KIMURA, K., TAKATA, S., FUJIKAWA, A., MORII, H., KUMASAKA, T., YAMAMOTO, M., ISHII, S. & OGATA, K. 2002. Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter. *Cell,* 108**,** 57-70.

TAKAGI, Y. & KORNBERG, R. D. 2006. Mediator as a general transcription factor. *Journal of Biological Chemistry,* 281**,** 80-89.

TAKENAKA, M., YAMAOKA, S., HANAJIRI, T., SHIMIZU-UEDA, Y., YAMATO, K. T., FUKUZAWA, H. & OHYAMA, K. 2000. Direct transformation and plant regeneration of the haploid liverwort Marchantia polymorpha L. *Transgenic Research,* 9**,** 179-185.

TAKEZAKI, N., RZHETSKY, A. & NEI, M. 1995. Phylogenetic Test of the Molecular Clock and Linearized Trees. *Molecular Biology and Evolution,* 12**,** 823-833.

TAMURA, K. & NEI, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol,* 10**,** 512-26.

TANAHASHI, T., SUMIKAWA, N., KATO, M. & HASEBE, M. 2005. Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss Physcomitrella patens. *Development,* 132**,** 1727-36.

TANG, H., BOWERS, J. E., WANG, X., MING, R., ALAM, M. & PATERSON, A. H. 2008a. Synteny and collinearity in plant genomes. *Science,* 320**,** 486-8.

TANG, H., BOWERS, J. E., WANG, X. & PATERSON, A. H. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A,* 107**,** 472-7.

TANG, H., WANG, X., BOWERS, J. E., MING, R., ALAM, M. & PATERSON, A. H. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res,* 18**,** 1944-54.

TANG, W., EZCURRA, I., MUSCHIETTI, J. & MCCORMICK, S. 2002. A cysteine-rich extracellular protein, LAT52, interacts with the extracellular domain of the pollen receptor kinase LePRK2. *Plant Cell,* 14**,** 2277-87.

THORNE, J. L., KISHINO, H. & PAINTER, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution,* 15**,** 1647-1657.

THORNE, R. F. & REVEAL, J. L. 2007. An updated classification of the class magnoliopsida ("Angiospermae"). *Botanical Review,* 73**,** 67-181.

TIMME, R. E., BACHVAROFF, T. R. & DELWICHE, C. F. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One,* 7**,** e29696.

TIMME, R. E. & DELWICHE, C. F. 2011. Phylogenomic Reconstruction of the Charophytes: A Multilocus Approach to Resolving the Phylogeny of Plants' Closest Relatives. *Journal of Phycology,* 47**,** 16-16.

TIMMERMANS, M. C. P. 1999. ROUGH SHEATH2: A Myb Protein That Represses knox Homeobox Genes in Maize Lateral Organ Primordia. *Science,* 284**,** 151-153.

TIWARI, S. B., BELACHEW, A., MA, S. F., YOUNG, M., ADE, J., SHEN, Y., MARION, C. M., HOLTAN, H. E., BAILEY, A., STONE, J. K., EDWARDS, L., WALLACE, A. D., CANALES, R. D., ADAM, L., RATCLIFFE, O. J. & REPETTI, P. P. 2012. The EDLL motif: a potent plant transcriptional activation domain from AP2/ERF transcription factors. *Plant J,* 70**,** 855-65.

TRIEZENBERG, S. J. 1995. Structure and function of transcriptional activation domains. *Curr Opin Genet Dev,* 5**,** 190-6.

TROUILLER, B., CHARLOT, F., CHOINARD, S., SCHAEFER, D. G. & NOGUE, F. 2007. Comparison of gene targeting efficiencies in two mosses suggests that it is a conserved feature of Bryophyte transformation. *Biotechnol Lett,* 29**,** 1591-8.

TROUILLER, B., SCHAEFER, D. G., CHARLOT, F. & NOGUE, F. 2006. MSH2 is essential for the preservation of genome integrity and prevents homeologous recombination in the moss Physcomitrella patens. *Nucleic Acids Res,* 34**,** 232-42.

TSUZUKI, M., NISHIHAMA, R., ISHIZAKI, K., KURIHARA, Y., MATSUI, M., BOWMAN, J. L., KOHCHI, T., HAMADA, T. & WATANABE, Y. 2016. Profiling and Characterization of Small RNAs in the Liverwort, Marchantia polymorpha, Belonging to the First Diverged Land Plants. *Plant Cell Physiol,* 57**,** 359-72.

TURMEL, M., OTIS, C. & LEMIEUX, C. 2006. The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. *Mol Biol Evol,* 23**,** 1324-38.

TURMEL, M., POMBERT, J. F., CHARLEBOIS, P., OTIS, C. & LEMIEUX, C. 2007. The green algal ancestry of land plants as revealed by the chloroplast genome. *International Journal of Plant Sciences,* 168**,** 679-689.

TWELL, D. 2011. Male gametogenesis and germline specification in flowering plants. *Sex Plant Reprod,* 24**,** 149-60.

TWELL, D., PARK, S. K. & LALANNE, E. 1998. Asymmetric division and cell-fate determination in developing pollen. *Trends in Plant Science,* 3**,** 305-310.

TWELL, D., YAMAGUCHI, J. & MCCORMICK, S. 1990. Pollen-specific gene expression in transgenic plants: coordinate regulation of two different tomato gene promoters during microsporogenesis. *Development,* 109**,** 705-13.

TWELL, D., YAMAGUCHI, J., WING, R. A., USHIBA, J. & MCCORMICK, S. 1991. Promoter Analysis of Genes That Are Coordinately Expressed during Pollen Development Reveals Pollen-Specific Enhancer Sequences and Shared Regulatory Elements. *Genes & Development,* 5**,** 496-507.

UEDA, M., TANAKA, A., SUGIMOTO, K., SHIKANAI, T. & NISHIMURA, Y. 2014. chlB requirement for chlorophyll biosynthesis under short photoperiod in Marchantia polymorpha L. *Genome Biol Evol,* 6**,** 620-8.

UESUGI, M., NYANGUILE, O., LU, H., LEVINE, A. J. & VERDINE, G. L. 1997. Induced alpha helix in the VP16 activation domain upon binding to a human TAF. *Science,* 277**,** 1310-1313.

VAN NIMWEGEN, E. 2003. Scaling laws in the functional content of genomes. *Trends Genet,* 19**,** 479-84.

VASUDEVAN, S., TONG, Y. C. & STEITZ, J. A. 2008. Cell cycle control of microRNA-mediated translation regulation. *Cell Cycle,* 7**,** 1545-1549.

VEITIA, R. A. 2002. Exploring the etiology of haploinsufficiency. *Bioessays,* 24**,** 175-84.

VEITIA, R. A. 2005. Gene dosage balance: deletions, duplications and dominance. *Trends Genet,* 21**,** 33-5.

VEITIA, R. A., BOTTANI, S. & BIRCHLER, J. A. 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet,* 24**,** 390-7.

VELMURUGU, Y. 2016. *Dynamics and mechanism of DNA-bending proteins in binding site recognition,* New York, NY, Springer Science+Business Media.

VERKEST, A., WEINL, C., INZE, D., DE VEYLDER, L. & SCHNITTGER, A. 2005. Switching the cell cycle. Kip-related proteins in plant cell cycle control. *Plant Physiol,* 139**,** 1099-106.

VESTY, E. F. 2010. *Conservation of DUO1 function in plant reproduction.* Thesis (M.Sc.), University of Leicester.

VINSON, C. R., SIGLER, P. B. & MCKNIGHT, S. L. 1989. Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science,* 246**,** 911-6.

VISION, T. J., BROWN, D. G. & TANKSLEY, S. D. 2000. The origins of genomic duplications in Arabidopsis. *Science,* 290**,** 2114-7.

VITTI, J. J. 2013. Cephalopod Cognition in an Evolutionary Context: Implications for Ethology. *Biosemiotics,* 6**,** 393-401.

VOINNET, O. 2009. Origin, biogenesis, and activity of plant microRNAs. *Cell,* 136**,** 669-87.

VOJNIC, E., MOURAO, A., SEIZL, M., SIMON, B., WENZECK, L., LARIVIERE, L., BAUMLI, S., BAUMGART, K., MEISTERERNST, M., SATTLER, M. & CRAMER, P. 2011. Structure and VP16 binding of the Mediator Med25 activator interaction domain. *Nat Struct Mol Biol,* 18**,** 404-9.

VON BESSER, K., FRANK, A. C., JOHNSON, M. A. & PREUSS, D. 2006. Arabidopsis HAP2 (GCS1) is a sperm-specific gene required for pollen tube guidance and fertilization. *Development,* 133**,** 4761-9.

WADA, T. 1997. Epidermal Cell Differentiation in Arabidopsis Determined by a Myb Homolog, CPC. *Science,* 277**,** 1113-1116.

WADA, T. 2002. Role of a positive regulator of root hair development, CAPRICE, in Arabidopsis root epidermal cell differentiation. *Development,* 129**,** 5409-5419.

WALBOT, V. & EVANS, M. M. 2003. Unique features of the plant life cycle and their consequences. *Nat Rev Genet,* 4**,** 369-79.

WANG, X., WANG, H., WANG, J., SUN, R., WU, J., LIU, S., BAI, Y., MUN, J. H., BANCROFT, I., CHENG, F., HUANG, S., LI, X., HUA, W., WANG, J., WANG, X., FREELING, M., PIRES, J. C., PATERSON, A. H., CHALHOUB, B., WANG, B., HAYWARD, A., SHARPE, A. G., PARK, B. S., WEISSHAAR, B., LIU, B., LI, B., LIU, B., TONG, C., SONG, C., DURAN, C., PENG, C., GENG, C., KOH, C., LIN, C., EDWARDS, D., MU, D., SHEN, D., SOUMPOUROU, E., LI, F., FRASER, F., CONANT, G., LASSALLE, G., KING, G. J., BONNEMA, G., TANG, H., WANG, H., BELCRAM, H., ZHOU, H., HIRAKAWA, H., ABE, H., GUO, H., WANG, H., JIN, H., PARKIN, I. A., BATLEY, J., KIM, J. S., JUST, J., LI, J., XU, J., DENG, J., KIM, J. A., LI, J., YU, J., MENG, J., WANG, J., MIN, J., POULAIN, J., WANG, J., HATAKEYAMA, K., WU, K., WANG, L., FANG, L., TRICK, M., LINKS, M. G., ZHAO, M., JIN, M., RAMCHIARY, N., DROU, N., BERKMAN, P. J., CAI, Q., HUANG, Q., LI, R., TABATA, S., CHENG, S., ZHANG, S., ZHANG, S., HUANG, S., SATO, S., SUN, S., KWON, S. J., CHOI, S. R., LEE, T. H., FAN, W., ZHAO, X., TAN, X., XU, X., WANG, Y., QIU, Y., YIN, Y., LI, Y., et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet,* 43**,** 1035-9.

WARNMARK, A., TREUTER, E., WRIGHT, A. P. & GUSTAFSSON, J. A. 2003. Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation. *Mol Endocrinol,* 17**,** 1901-9.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WELLMAN, C. H., OSTERLOFF, P. L. & MOHIUDDIN, U. 2003. Fragments of the earliest land plants. *Nature,* 425**,** 282-5.

WESTON, K. 1998. Myb proteins in life, death and differentiation. *Curr Opin Genet Dev,* 8**,** 76-81.

WICKETT, N. J., MIRARAB, S., NGUYEN, N., WARNOW, T., CARPENTER, E., MATASCI, N., AYYAMPALAYAM, S., BARKER, M. S., BURLEIGH, J. G., GITZENDANNER, M. A., RUHFEL, B. R., WAFULA, E., DER, J. P., GRAHAM, S. W., MATHEWS, S., MELKONIAN, M., SOLTIS,

D. E., SOLTIS, P. S., MILES, N. W., ROTHFELS, C. J., POKORNY, L., SHAW, A. J., DEGIRONIMO, L., STEVENSON, D. W., SUREK, B., VILLARREAL, J. C., ROURE, B., PHILIPPE, H., DEPAMPHILIS, C. W., CHEN, T., DEYHOLOS, M. K., BAUCOM, R. S., KUTCHAN, T. M., AUGUSTIN, M. M., WANG, J., ZHANG, Y., TIAN, Z., YAN, Z., WU, X., SUN, X., WONG, G. K. & LEEBENS-MACK, J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A,* 111**,** E4859-68.

WIEDERSTEIN, M. & SIPPL, M. J. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res,* 35**,** W407-10.

WIGHTMAN, B., BURGLIN, T. R., GATTO, J., ARASU, P. & RUVKUN, G. 1991. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes & Development,* 5**,** 1813-1824.

WIGHTMAN, B., HA, I. & RUVKUN, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-62.

WILBURN, D. B. & SWANSON, W. J. 2015. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *J Proteomics*.

WILKINS, M. R., GASTEIGER, E., BAIROCH, A., SANCHEZ, J. C., WILLIAMS, K. L., APPEL, R. D. & HOCHSTRASSER, D. F. 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol,* 112**,** 531-52.

WILSON, Z. A., MORROLL, S. M., DAWSON, J., SWARUP, R. & TIGHE, P. J. 2001. The Arabidopsis MALE STERILITY1 (MS1) gene is a transcriptional regulator of male gametogenesis, with homology to the PHD-finger family of transcription factors. *Plant Journal,* 28**,** 27-39.

WILSON, Z. A. & YANG, C. Y. 2004. Plant gametogenesis: conservation and contrasts in development. *Reproduction,* 128**,** 483-492.

WINTER, D., VINEGAR, B., NAHAL, H., AMMAR, R., WILSON, G. V. & PROVART, N. J. 2007. An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One,* 2**,** e718.

WODNIOK, S., BRINKMANN, H., GLOCKNER, G., HEIDEL, A. J., PHILIPPE, H., MELKONIAN, M. & BECKER, B. 2011. Origin of land plants: do conjugating green algae hold the key? *BMC Evol Biol,* 11**,** 104.

WOLFE, S. A., NEKLUDOVA, L. & PABO, C. O. 2000. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct,* 29**,** 183-212.

WONG, J. L. & JOHNSON, M. A. 2010. Is HAP2-GCS1 an ancestral gamete fusogen? *Trends Cell Biol,* 20**,** 134-41.

WOOD, A. J., OLIVER, M. J. & COVE, D. J. 2000. Bryophytes as model systems. *Bryologist,* 103**,** 128-133.

WYSOCKA, J. & HERR, W. 2003. The herpes simplex virus VP16-induced complex: the makings of a regulatory switch. *Trends in Biochemical Sciences,* 28**,** 294-304.

XU, J., DING, Z. W., VIZCAY-BARRENA, G., SHI, J. X., LIANG, W. Q., YUAN, Z., WERCK-REICHHART, D., SCHREIBER, L., WILSON, Z. A. & ZHANG, D. B. 2014a. ABORTED MICROSPORES Acts as a Master Regulator of Pollen Wall Formation in Arabidopsis. *Plant Cell,* 26**,** 1544-1556.

XU, J., YANG, C. Y., YUAN, Z., ZHANG, D. S., GONDWE, M. Y., DING, Z. W., LIANG, W. Q., ZHANG, D. B. & WILSON, Z. A. 2010. The ABORTED MICROSPORES Regulatory Network Is Required for Postmeiotic Male Reproductive Development in Arabidopsis thaliana. *Plant Cell,* 22**,** 91-107.

XU, P., VERNOOY, S. Y., GUO, M. & HAY, B. A. 2003. The Drosophila MicroRNA Mir-14 Suppresses Cell Death and Is Required for Normal Fat Metabolism. *Current Biology,* 13**,** 790-795.

XU, Y., IACUONE, S., LI, S. F. & PARISH, R. W. 2014b. MYB80 homologues in Arabidopsis, cotton and Brassica: regulation and functional conservation in tapetal and pollen development. *BMC Plant Biol,* 14**,** 278.

YADEGARI, R. & DREWS, G. N. 2004. Female gametophyte development. *Plant Cell,* 16**,** S133-S141.

YAMAWAKI, S., YAMASHINO, T., NAKAMICHI, N., NAKANISHI, H. & MIZUNO, T. 2011. Light-responsive double B-box containing transcription factors are conserved in Physcomitrella patens. *Biosci Biotechnol Biochem,* 75**,** 2037-41.

YANAGISAWA, S. 2001. The transcriptional activation domain of the plant-specific Dof1 factor functions in plant, animal, and yeast cells. *Plant Cell Physiol,* 42**,** 813-22.

YANG, C., VIZCAY-BARRENA, G., CONNER, K. & WILSON, Z. A. 2007a. MALE STERILITY1 is required for tapetal development and pollen wall biosynthesis. *Plant Cell,* 19**,** 3530-48.

YANG, C. Y., XU, Z. Y., SONG, J., CONNER, K., BARRENA, G. V. & WILSON, Z. A. 2007b. Arabidopsis MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence. *Plant Cell,* 19**,** 534-548.

YANG, W. C., SHI, D. Q. & CHEN, Y. H. 2010. Female gametophyte development in flowering plants. *Annu Rev Plant Biol,* 61**,** 89-108.

YANG, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol,* 15**,** 568-73.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YANG, Z. & DOS REIS, M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol,* 28**,** 1217-28.

YANG, Z. H. & NIELSEN, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution,* 19**,** 908-917.

YANG, Z. H., NIELSEN, R., GOLDMAN, N. & PEDERSEN, A. M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 155**,** 431-449.

YANHUI, C., XIAOYUAN, Y., KUN, H., MEIHUA, L., JIGANG, L., ZHAOFENG, G., ZHIQIANG, L., YUNFEI, Z., XIAOXIAO, W., XIAOMING, Q., YUNPING, S., LI, Z., XIAOHUI, D., JINGCHU, L., XING-WANG, D., ZHANGLIANG, C., HONGYA, G. & LI-JIA, Q. 2006. The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol,* 60**,** 107-24.

YEKTA, S., SHIH, I. H. & BARTEL, D. P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science,* 304**,** 594-6.

YU, E. Y., KIM, S. E., KIM, J. H., KO, J. H., CHO, M. H. & CHUNG, I. K. 2000. Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. *J Biol Chem,* 275**,** 24208-14.

YU, M. & IRWIN, D. M. 1996. Evolution of stomach lysozyme: the pig lysozyme gene. *Mol Phylogenet Evol,* 5**,** 298-308.

ZHANG, B., PAN, X. & STELLWAG, E. J. 2008. Identification of soybean microRNAs and their targets. *Planta,* 229**,** 161-82.

ZHANG, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol,* 21**,** 1332-9.

ZHANG, J., KUMAR, S. & NEI, M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol,* 14**,** 1335-8.

ZHANG, J., NIELSEN, R. & YANG, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol,* 22**,** 2472-9.

ZHANG, J., ROSENBERG, H. F. & NEI, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A,* 95**,** 3708-13.

ZHANG, S., XIE, M., REN, G. & YU, B. 2013. CDC5, a DNA binding protein, positively regulates posttranscriptional processing and/or transcription of primary microRNA transcripts. *Proc Natl Acad Sci U S A,* 110**,** 17588-93.

ZHAO, T., LI, G., MI, S., LI, S., HANNON, G. J., WANG, X. J. & QI, Y. 2007. A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev,* 21**,** 1190-203.

ZHENG, B., HE, H., ZHENG, Y., WU, W. & MCCORMICK, S. 2014. An ARID domain-containing protein within nuclear bodies is required for sperm cell formation in Arabidopsis thaliana. *PLoS Genet,* 10**,** e1004421.

ZHONG, B., LIU, L., YAN, Z. & PENNY, D. 2013. Origin of land plants using the multispecies coalescent model. *Trends Plant Sci,* 18**,** 492-5.

ZHU, J., CHEN, H., LI, H., GAO, J. F., JIANG, H., WANG, C., GUAN, Y. F. & YANG, Z. N. 2008. Defective in Tapetal development and function 1 is essential for anther development and tapetal function for microspore maturation in Arabidopsis. *Plant J,* 55**,** 266-77.

ZHU, J., ZHANG, G. Q., CHANG, Y. H., LI, X. C., YANG, J., HUANG, X. Y., YU, Q. B., CHEN, H., WU, T. L. & YANG, Z. N. 2010. AtMYB103 is a crucial regulator of several pathways affecting Arabidopsis anther development. *Science China-Life Sciences,* 53**,** 1112-1122.

ZUO, J., NIU, Q. W. & CHUA, N. H. 2000. Technical advance: An estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants. *Plant J,* 24**,** 265-73.